# EXPRESSION PROFILES FOR BREAST CANCER AND METHODS OF USE

**[001]** This application claims benefit of U.S. Provisional Application Serial No. 60/450,655, February 28, 2003, the contents of which are incorporated herein by reference in their entirety.

## FIELD OF THE INVENTION

**[002]** The present invention relates to gene expression profiles for breast cancer, microarrays comprising nucleic acid sequences representing gene expression profiles, and methods of using expression profiles and microarrays.

## BACKGROUND OF THE INVENTION

**[003]** Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription of particular genes (e.g., through control of initiation, provision of RNA precursors, RNA processing, etc.). For example, losses and gains of genetic material play an important role in malignant transformation and progression. These gains and losses are thought to be "driven" by at least two kinds of genes, oncogenes and tumor suppressor genes. Oncogenes are positive regulators of tumorgenesis, while tumor suppressor genes are negative regulators of tumorgenesis (Marshall, Cell 64:313-326, 1991; Weinberg, Science 254:1138-1146, 1991). Therefore, one mechanism of activating unregulated growth is to increase the number of genes coding for oncogene proteins or to increase the level of expression of these oncogenes (e.g., in response to cellular or environmental changes), and another mechanism is to lose genetic material or to decrease the level of expression of genes that code for tumor suppressors. This model is supported by the losses and gains of genetic material associated with glioma progression (Mikkelson, et al., J. Cellular Biochem. 46:3-8, 1991). Thus, changes in the expression (transcription) levels of particular genes (e.g., oncogenes or tumor suppressors) serve as signposts for the presence and progression of various cancers.

**[004]** Compounds which are used as therapeutics to treat these various diseases (e.g., cancer) presumably reverse some, or all, of these gene expression changes. The expression change of at least some of these genes may therefore, be used as a method to monitor, or even predict, the efficacy of such therapeutics. The analysis of these expression changes may be performed in the target tissue of interest (e.g., tumor) or in some surrogate cell population (e.g., peripheral blood leukocytes). In the latter case, correlation of the gene expression changes with efficacy (e.g., tumor shrinkage or non-growth) must be especially strong for the expression change pattern to be used as a marker for efficacy.

Date of Deposit: February 27, 2004
Express Mail Label No. EL 844820261 US

[005] A number of laboratories have reported success in using gene expression analysis, via microarrays or other methods, to classify human tumors at the molecular level (Bittner, et al., Nature 406:536-540, 2000; Alon, et al., Proc. Natl. Acad. Sci. USA 96:6745-6750, 1999; Alizadeh, et al., Nature 403:503-511, 2000; Golub, et al., Science 286:531-537, 1999; Perou, et al., Proc. Natl. Acad. Sci. 96:9212-9217, 1999; Kahn, et al., Am. J. Pathol. 156:1887-1900, 2000). Genes, either individually or as a subset, identified in this way may be used as markers that could be tracked for changes that correlate with efficacy of a therapeutic compound(s) or to predict which patients might benefit from a particular therapeutic. Total RNA was isolated from ten human breast tumors and from normal adjacent tissue (NAT), and the RNA was analyzed from each sample using Affymetrix technology.

## SUMMARY OF THE INVENTION

[006] The present invention relates to gene expression profiles for breast cancer, microarrays comprising nucleic acid sequences or amino acid sequences representing expression profiles, and methods of using expression profiles and microarrays.

[007] In one embodiment of the present invention, the gene expression profile is an expression profile comprising one or more genes (e.g., SEQ ID NOs: 1-127) that demonstrate altered expression in human breast tumors versus normal adjacent tissue (NAT).

[008] In another embodiment, the expression profile is an expression profile comprising one or more polypeptides (e.g., SEQ ID NOs: 128-254) that demonstrate altered expression in human breast tumors versus normal adjacent tissue (NAT).

[009] In further embodiment of the present invention, the gene expression profile may be an expression profile comprising one or more genes selected from the group consisting of the genes listed in the Table 1 to 3. In another embodiment of the present invention, the gene expression profiles comprise one or more biomarkers isolated from the group comprising the genes listed in the Tables.

[010] The present invention is also directed to the discovery of the gene expression profile of human breast tumors and normal adjacent tissue. As described in the Examples and in the Tables, human breast tumors have genes which are expressed at higher levels (i.e., which are up-regulated) and genes which are expressed at lower levels (i.e., which are down-regulated) relative to normal adjacent tissue. Sets of genes which are up-regulated or down-regulated are referred to herein as "genes characteristic of human breast tumor tissue."

[011] Also within the scope of the present invention are microarrays comprising one or more genes that demonstrate altered expression in human breast tumor tissue. In another embodiment of

the present invention, the microarray may be a microarray comprising one or more genes selected from the group consisting of the genes listed in the Tables. In a further embodiment, the microarray may be a microarray comprising one or more biomarkers isolated from the group comprising the genes listed in the Tables.

**[012]** In addition, it is an objective of the invention to provide methods and reagents for the prediction, diagnosis, prognosis, and therapy of cancer.

**[013]** This invention also relates to methods for using said microarrays which include, but are not limited to, screening the effects of a drug or treatment on tissue or cell samples, screening toxicity effects on tissue or cell samples, identifying a disease state in a tissue or cell sample, providing a patient diagnosis, predicting a patient's response to treatment, distinguishing between control and drug-treated samples, distinguishing between normal and tumor samples, discovering novel drugs, and determining the level of gene expression in a tissue or cell sample.

**[014]** Another embodiment of the present invention is a method for screening the effects of a drug on a tissue or cell sample comprising the step of analyzing the level of expression of one or more genes (e.g., SEQ ID NOs: 1-127) and/or gene products (e.g., SEQ ID NOs: 128-254), wherein the gene expression and/or gene product levels in the tissue or cell sample are analyzed before and after exposure to the drug, and a variation in the expression level of the gene and/or gene product is indicative of a drug effect or provides a patient diagnosis or predicts a patient's response to the treatment.

**[015]** Another aspect of the present invention is a method for discovering novel drugs comprising the step of analyzing the level of expression of one or more genes and/or gene products, wherein the gene expression and/or gene product levels of the cells are analyzed before and after exposure to the drug, and a variation in the expression level of the gene and/or gene product is indicative of drug efficacy.

**[016]** The invention further provides a method for identifying a compound useful for the treatment of cancer comprising administering to a subject with cancer a test compound, and measuring the activity of the polypeptide (e.g., the polypeptides encoded by SEQ ID NOs: 128-254), wherein a change in the activity of the polypeptide is indicative of the test compound being useful for the treatment of cancer.

**[017]** The invention, thus, provides methods which may be used to identify compounds which may act, for example, as regulators or modulators such as agonists and antagonists, partial agonists, inverse agonists, activators, co-activators, and inhibitors. Accordingly, the invention provides reagents and methods for regulating the expression of a polynucleotide or a polypeptide associated with cancer. Reagents that modulate the expression, stability, or amount of a

polynucleotide or the activity of the polypeptide may be a protein, a peptide, a peptidomimetic, a nucleic acid, a nucleic acid analogue (e.g., peptide nucleic acid, locked nucleic acid), or a small molecule.

**[018]** The present invention also provides a method for providing a patient diagnosis comprising the step of analyzing the level of expression of one or more genes and/or gene products, wherein the gene expression and/or gene product levels of normal and patient samples are analyzed, and a variation in the expression level of the gene and/or gene product in the patient sample is diagnostic of a disease. The patient samples include, but are not limited to, blood, amniotic fluid, plasma, semen, bone marrow, and tissue biopsy.

**[019]** The present invention still further provides a method of diagnosing cancer in a subject comprising measuring the activity of the polypeptide in a subject suspected of having cancer, wherein if there is a difference in the activity of the polypeptide, relative to the activity of the polypeptide in a subject not suspected of having cancer, then the subject is diagnosed has having cancer.

**[020]** In another embodiment, the invention provides a method for detecting cancer in a patient sample in which an antibody to a protein is used to react with proteins in the patient sample.

**[021]** Another aspect of the present invention is a method for distinguishing between normal and disease states comprising the step of analyzing the level of expression of one or more genes and/or gene products, wherein the gene expression and/or gene product levels of normal and disease tissues are analyzed, and a variation in the expression level of the gene and/or gene product is indicative of a disease state.

**[022]** In another embodiment, the invention pertains to a method of determining the phenotype of cells comprising detecting the differential expression, relative to normal cells, of at least one gene, wherein the gene is differentially expressed by at least a factor of two, at least a factor of five, at least a factor of twenty, or at least a factor of fifty.

**[023]** In yet another embodiment, the invention pertains to a method of determining the phenotype of cells, comprising detecting the differential expression, relative to normal cells, of at least one polypeptide, wherein the protein is differentially expressed by at least a factor of two, at least a factor of five, at least a factor of twenty, an up to at least a factor of fifty.

**[024]** In another embodiment, the invention pertains to a method for determining the phenotype of cells from a patient by providing a nucleic acid probe comprising a nucleotide sequence having at least about 10, at least about 15, at least about 25, or at least about 40 consecutive nucleotides, obtaining a sample of cells from a patient, optionally providing a second sample of cells substantially all of which are non-cancerous, contacting the nucleic acid probe under stringent

conditions with mRNA of each of said first and second cell samples, and comparing (a) the amount of hybridization of the probe with mRNA of the first cell sample, with (b) the amount of hybridization of the probe with mRNA of the second cell sample, wherein a difference of at least a factor of two, at least a factor of five, at least a factor of twenty, or at least a factor of fifty in the amount of hybridization with the mRNA of the first cell sample as compared to the amount of hybridization with the mRNA of the second cell sample is indicative of the phenotype of cells in the first cell sample.

**[025]** In another embodiment, the invention provides a test kit for identifying the presence of cancerous cells or tissues, comprising a probe/primer, for measuring a level of a nucleic acid in a sample of cells isolated from a patient. In certain embodiments, the kit may further include instructions for using the kit, solutions for suspending or fixing the cells, detectable tags or labels, solutions for rendering a nucleic acid susceptible to hybridization, solutions for lysing cells, or solutions for the purification of nucleic acids.

**[026]** In one embodiment, the invention provides a test kit for identifying the presence of cancer cells or tissues, comprising an antibody specific for a protein. In certain embodiments, the kit further includes instructions for using the kit. In certain embodiments, the kit may further include solutions for suspending or fixing the cells, detectable tags or labels, solutions for rendering a polypeptide susceptible to the binding of an antibody, solutions for lysing cells, or solutions for the purification of polypeptides.

**[027]** In another embodiment, the invention provides a test kit for monitoring the efficacy of a compound or therapeutic in cancerous cells or tissues, comprising a probe/primer, for measuring a level of a nucleic acid in a sample of cells isolated from a patient. In certain embodiments, the kit may further include instructions for using the kit, solutions for suspending or fixing the cells, detectable tags or labels, solutions for rendering a nucleic acid susceptible to hybridization, solutions for lysing cells, or solutions for the purification of nucleic acids.

**[028]** In one embodiment, the invention provides a test kit for monitoring the efficacy of a compound or therapeutic in cancer cells or tissues, comprising an antibody specific for a protein. In certain embodiments, the kit further includes instructions for using the kit. In certain embodiments, the kit may further include solutions for suspending or fixing the cells, detectable tags or labels, solutions for rendering a polypeptide susceptible to the binding of an antibody, solutions for lysing cells, or solutions for the purification of polypeptides.

**[029]** This invention is also related to methods of identifying biomarkers comprising the steps of selecting a set of biomarker genes from a gene expression profile representing a disease or drug treatment

## DETAILED DESCRIPTION OF THE INVENTION

**[030]** It is to be understood that this invention is not limited to the particular methodology, protocols, cell lines, animal species or genera, constructs, and reagents described and as such may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

**[031]** It must be noted that as used herein and in the appended claims, the singular forms "a," "and," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, reference to "a gene" is a reference to one or more genes and includes equivalents thereof known to those skilled in the art, and so forth.

**[032]** Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs. Any methods, devices, and materials similar or equivalent to those described herein can be used in the practice or testing of the invention, and examples of such methods, devices and materials are described below.

**[033]** All publications and patents mentioned herein are hereby incorporated herein by reference for the purpose of describing and disclosing, for example, the constructs and methodologies that are described in the publications which might be used in connection with the presently described invention. The publications discussed above and throughout the text are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention.

**Definitions**

**[034]** For convenience, the meaning of certain terms and phrases employed in the specification, examples, and appended claims are provided below.

**[035]** The phrase "a corresponding normal cell of" or "normal cell corresponding to" or "normal counterpart cell of" a diseased cell refers to a normal cell of the same type as that of the diseased cell.

**[036]** An "address" on an array (e.g., a microarray) refers to a location at which an element, for example, an oligonucleotide, is attached to the solid surface of the array.

**[037]** The term "agonist," as used herein, is meant to refer to an agent that mimics or up-regulates (e.g., potentiates or supplements) the bioactivity of a protein. An agonist may be a wild-type protein or derivative thereof having at least one bioactivity of the wild-type protein. An agonist

may also be a compound that up-regulates expression of a gene or which increases at least one bioactivity of a protein. An agonist can also be a compound which increases the interaction of a polypeptide with another molecule, for example, a target peptide or nucleic acid.

**[038]** "Amplification," as used herein, relates to the production of additional copies of a nucleic acid sequence. For example, amplification may be carried out using polymerase chain reaction (PCR) technologies which are well known in the art. (*see, e.g.,* Dieffenbach, C. W. and G. S. Dveksler (1995) PCR Primer, A Laboratory Manual, Cold Spring Harbor Press, Plainview, N.Y.)

**[039]** "Antagonist," as used herein, is meant to refer to an agent that down-regulates (e.g., suppresses or inhibits) at least one bioactivity of a protein. An antagonist may be a compound which inhibits or decreases the interaction between a protein and another molecule, for example, a target peptide or enzyme substrate. An antagonist may also be a compound that down-regulates expression of a gene or which reduces the amount of expressed protein present.

**[040]** The term "antibody," as used herein, is intended to include whole antibodies, for example, of any isotype (IgG, IgA, IgM, IgE, etc.), and includes fragments thereof which are also specifically reactive with a vertebrate (e.g., mammalian) protein. Antibodies may be fragmented using conventional techniques and the fragments screened for utility in the same manner as described above for whole antibodies. Thus, the term includes segments of proteolytically-cleaved or recombinantly-prepared portions of an antibody molecule that are capable of selectively reacting with a certain protein. Non-limiting examples of such proteolytic and/or recombinant fragments include Fab, F(ab')2, Fab', Fv, and single chain antibodies (scFv) containing a V[L] and/or V[H] domain joined by a peptide linker. The scFv's may be covalently or non-covalently linked to form antibodies having two or more binding sites. The subject invention includes polyclonal, monoclonal, or other purified preparations of antibodies and recombinant antibodies.

**[041]** The terms "array" or "matrix" refer to an arrangement of addressable locations or "addresses" on a device. The locations can be arranged in two-dimensional arrays, three-dimensional arrays, or other matrix formats. The number of locations may range from several to at least hundreds of thousands. Most importantly, each location represents a totally independent reaction site. A "nucleic acid array" refers to an array containing nucleic acid probes, such as oligonucleotides or larger portions of genes. The nucleic acid on the array may be single-stranded. Arrays wherein the probes are oligonucleotides are referred to as "oligonucleotide arrays" or "oligonucleotide chips." A "microarray," also referred to herein as a "biochip" or "biological chip," is an array of regions having a density of discrete regions of, for example, at least about $100/cm^2$, or at least about $1000/cm^2$. The regions in a microarray have typical dimensions, for example, diameters, in the range of between about 10-250 μm, and are separated from other regions in the array by about the same distance.

[042] "Biological activity," "bioactivity," "activity," or "biological function," which are used interchangeably, herein mean an effector or antigenic function that is directly or indirectly performed by a polypeptide (whether in its native or denatured conformation), or by any subsequence thereof. Biological activities include binding to polypeptides, binding to other proteins or molecules, activity as a DNA binding protein, as a transcription regulator, ability to bind damaged DNA, etc. A bioactivity can be modulated by directly affecting the subject polypeptide. Alternatively, a bioactivity can be altered by modulating the level of the polypeptide, such as by modulating expression of the corresponding gene.

[043] The term "biological sample," as used herein, refers to a sample obtained from an organism or from components (e.g., cells) of an organism. The sample may be of any biological tissue or fluid. The sample may be a "clinical sample" which is a sample derived from a patient. Such samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

[044] The term "biomarker" or "marker" encompasses a broad range of intra- and extra-cellular events as well as whole-organism physiological changes. Biomarkers may be represent essentially any aspect of cell function, for example, but not limited to, levels or rate of production of signaling molecules, transcription factors, metabolites, gene transcripts as well as post-translational modifications of proteins. Biomarkers may include whole genome analysis of transcript levels or whole proteome analysis of protein levels and/or modifications.

[045] A biomarker may also refer to a gene or gene product which is up- or down-regulated in a compound-treated, diseased cell of a subject having the disease compared to an untreated diseased cell. That is, the gene or gene product is sufficiently specific to the treated cell that it may be used, optionally with other genes or gene products, to identify, predict, or detect efficacy of a small molecule. Thus, a biomarker is a gene or gene product that is characteristic of efficacy of a compound in a diseased cell or the response of that diseased cell to treatment by the compound.

[046] A nucleotide sequence is "complementary" to another nucleotide sequence if each of the bases of the two sequences match, that is, are capable of forming Watson-Crick base pairs. The term "complementary strand" is used herein interchangeably with the term "complement." The complement of a nucleic acid strand may be the complement of a coding strand or the complement of a non-coding strand.

[047] "Detection agents of genes" refers to agents that can be used to specifically detect the gene or other biological molecules relating to it, for example, RNA transcribed from the gene or

polypeptides encoded by the gene. Exemplary detection agents are nucleic acid probes, which hybridize to nucleic acids corresponding to the gene, and antibodies.

**[048]** "Differential gene expression pattern" between cell A and cell B refers to a pattern reflecting the differences in gene expression between cell A and cell B. A differential gene expression pattern may also be obtained between a cell at one time point and a cell at another time point, or between a cell incubated or contacted with a compound and a cell that has not been incubated with or contacted with the compound.

**[049]** The term "cancer" includes, but is not limited to, solid tumors, such as cancers of the breast, respiratory tract, brain, reproductive organs, digestive tract, urinary tract, eye, liver, skin, head and neck, thyroid, parathyroid, and their distant metastases. The term also includes lymphomas, sarcomas, and leukemias.

**[050]** Examples of breast cancer include, but are not limited to, invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma *in situ*, and lobular carcinoma *in situ.*

**[051]** Examples of cancers of the respiratory tract include, but are not limited to, small-cell and non-small-cell lung carcinoma, as well as bronchial adenoma and pleuropulmonary blastoma.

**[052]** Examples of brain cancers include, but are not limited to, brain stem and hypophtalmic glioma, cerebellar and cerebral astrocytoma, medulloblastoma, ependymoma, as well as neuroectodermal and pineal tumor.

**[053]** Tumors of the male reproductive organs include, but are not limited to, prostate and testicular cancer. Tumors of the female reproductive organs include, but are not limited to, endometrial, cervical, ovarian, vaginal, and vulvar cancer, as well as sarcoma of the uterus.

**[054]** Tumors of the digestive tract include, but are not limited to, anal, colon, colorectal, esophageal, gallbladder, gastric, pancreatic, rectal, small-intestine, and salivary gland cancers.

**[055]** Tumors of the urinary tract include, but are not limited to, bladder, penile, kidney, renal pelvis, ureter, and urethral cancers.

**[056]** Eye cancers include, but are not limited to, intraocular melanoma and retinoblastoma.

**[057]** Examples of liver cancers include, but are not limited to, hepatocellular carcinoma (liver cell carcinomas with or without fibrolamellar variant), cholangiocarcinoma (intrahepatic bile duct carcinoma), and mixed hepatocellular cholangiocarcinoma.

**[058]** Skin cancers include, but are not limited to, squamous cell carcinoma, Kaposi's sarcoma, malignant melanoma, Merkel cell skin cancer, and non-melanoma skin cancer.

**[059]** Head-and-neck cancers include, but are not limited to, laryngeal / hypopharyngeal / nasopharyngeal / oropharyngeal cancer, and lip and oral cavity cancer.

**[060]** Lymphomas include, but are not limited to, AIDS-related lymphoma, non-Hodgkin's lymphoma, cutaneous T-cell lymphoma, Hodgkin's disease, and lymphoma of the central nervous system.

**[061]** Sarcomas include, but are not limited to, sarcoma of the soft tissue, osteosarcoma, malignant fibrous histiocytoma, lymphosarcoma, and rhabdomyosarcoma.

**[062]** Leukemias include, but are not limited to, acute myeloid leukemia, acute lymphoblastic leukemia, chronic lymphocytic leukemia, chronic myelogenous leukemia, and hairy cell leukemia.

**[063]** "A diseased cell of cancer" refers to a cell present in subjects having cancer. That is, a cell which is a modified form of a normal cell and is not present in a subject not having cancer, or a cell which is present in significantly higher or lower numbers in subjects having cancer relative to subjects not having cancer.

**[064]** The term "equivalent" is understood to include nucleotide sequences encoding functionally equivalent polypeptides. Equivalent nucleotide sequences may include sequences that differ by one or more nucleotide substitutions, additions, or deletions, such as allelic variants; and may, therefore, include sequences that differ from the nucleotide sequence of the nucleic acids referred to in the Tables due to the degeneracy of the genetic code.

**[065]** The term "expression profile," which is used interchangeably herein with "gene expression profile" and "fingerprint" of a cell refers to a set of values representing mRNA levels of one or more genes in a cell. An expression profile may comprise values representing expression levels of, for example, at least about 10 genes, or at least about 50, 100, 200 or more genes. Expression profiles may also comprise an mRNA level of a gene which is expressed at similar levels in multiple cells and conditions (e.g., a housekeeping gene such as GAPDH). For example, an expression profile of a diseased cell of cancer refers to a set of values representing mRNA levels of 10 or more genes in a diseased cell. In addition, the term "expression profile" may also include a set of values representing one or more protein or polypeptide levels in a cell.

**[066]** The term "gene" refers to a nucleic acid sequence that comprises control and coding sequences necessary for the production of a polypeptide or precursor. The polypeptide can be encoded by a full length coding sequence or by any portion of the coding sequence. The gene may be derived in whole or in part from any source known to the art, including a plant, a fungus, an animal, a bacterial genome or episome, eukaryotic, nuclear or plasmid DNA, cDNA, viral DNA, or chemically synthesized DNA. A gene may contain one or more modifications in either the coding or the untranslated regions which could affect the biological activity or the chemical structure of

the expression product, the rate of expression, or the manner of expression control. Such modifications include, but are not limited to, mutations, insertions, deletions, and substitutions of one or more nucleotides. The gene may constitute an uninterrupted coding sequence or it may include one or more introns, bound by the appropriate splice junctions.

**[067]** "Hybridization" refers to any process by which a strand of nucleic acid binds with a complementary strand through base pairing. For example, two single-stranded nucleic acids "hybridize" when they form a double-stranded duplex. The region of double-strandedness may include the full-length of one or both of the single-stranded nucleic acids, or all of one single-stranded nucleic acid and a subsequence of the other single-stranded nucleic acid, or the region of double-strandedness may include a subsequence of each nucleic acid. Hybridization also includes the formation of duplexes which contain certain mismatches, provided that the two strands are still forming a double-stranded helix. "Stringent hybridization conditions" refers to hybridization conditions resulting in essentially specific hybridization.

**[068]** The term "isolated," as used herein, with respect to nucleic acids, such as DNA or RNA, refers to molecules separated from other DNAs or RNAs, respectively, that are present in the natural source of the macromolecule. The term "isolated" as used herein also refers to a nucleic acid or peptide that is substantially free of cellular material, viral material, culture medium when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Moreover, an "isolated nucleic acid" may include nucleic acid fragments which are not naturally occurring as fragments and would not be found in the natural state. The term "isolated" is also used herein to refer to polypeptides which are isolated from other cellular proteins and is meant to encompass both purified and recombinant polypeptides.

**[069]** As used herein, the terms "label" and "detectable label" refer to a molecule capable of detection, including, but not limited to, radioactive isotopes, fluorophores, chemiluminescent moieties, enzymes, enzyme substrates, enzyme cofactors, enzyme inhibitors, dyes, metal ions, ligands (e.g., biotin or haptens), and the like. The term "fluorescer" refers to a substance or a portion thereof which is capable of exhibiting fluorescence in the detectable range. Particular examples of labels which may be used in the present invention include fluorescein, rhodamine, dansyl, umbelliferone, Texas red, luminol, NADPH, alpha - beta –galactosidase, and horseradish peroxidase.

**[070]** The phrase "level of expression" refers to the level of mRNA, as well as pre-mRNA nascent transcript(s), transcript processing intermediates, mature mRNA(s), and degradation products, encoded by a gene in the cell. The phrase "level of expression" also refers to the level of protein or polypeptide in a cell.

**[071]** As used herein, the term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA) and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as equivalents, analogs of either RNA or DNA made from nucleotide analogs and, as applicable to the embodiment being described, single-stranded (sense or antisense) and double-stranded polynucleotides. Chromosomes, cDNAs, mRNAs, rRNAs, and ESTs are representative examples of molecules that may be referred to as nucleic acids.

**[072]** The phrase "nucleic acid corresponding to a gene" refers to a nucleic acid that can be used for detecting the gene, for example, a nucleic acid which is capable of hybridizing specifically to the gene.

**[073]** The phrase "nucleic acid sample derived from RNA" refers to one or more nucleic acid molecules (e.g., RNA or DNA) that may be synthesized from the RNA, and includes DNA produced from methods using PCR (e.g., RT-PCR).

**[074]** The term "oligonucleotide" as used herein refers to a nucleic acid molecule comprising, for example, from about 10 to about 1000 nucleotides. Oligonucleotides for use in the present invention may be, for example, from about 15 to about 150 nucleotides, or from about 150 to about 1000 in length. The oligonucleotide may be a naturally occurring oligonucleotide or a synthetic oligonucleotide. Oligonucleotides may be prepared by the phosphoramidite method (Beaucage and Carruthers, Tetrahedron Lett. 22:1859-62, 1981), or by the triester method (Matteucci, et al., J. Am. Chem. Soc. 103:3185, 1981), or by other chemical methods known in the art.

**[075]** The term "patient" or "subject" as used herein includes mammals (e.g., humans and animals).

**[076]** The term "percent identical" refers to sequence identity between two amino acid sequences or between two nucleotide sequences. For example, identity between two sequences may be determined by comparing a particular position in each sequence which may be aligned for purposes of comparison. When an equivalent position in the compared sequences is occupied by the same base or amino acid, then the molecules are identical at that position. When the equivalent site is occupied by the same or a similar amino acid residue (e.g., similar in steric and/or electronic nature), then the molecules may be referred to as homologous (similar) at that position. Expression as a percentage of homology, similarity, or identity refers to a function of the number of identical or similar amino acids at positions shared by the compared sequences. Various alignment algorithms and/or programs may be used including, for example, FASTA, BLAST, or ENTREZ. FASTA and BLAST are available as a part of the GCG sequence analysis package (University of Wisconsin, Madison, Wis.), and may be used with, for example, default settings. ENTREZ is available through the National Center for Biotechnology Information, National

Library of Medicine, National Institutes of Health, Bethesda, MD. In one embodiment, the percent identity of two sequences may be determined by the GCG program with a gap weight of 1 (e.g., each amino acid gap is weighted as if it were a single amino acid or nucleotide mismatch between the two sequences). Other techniques for alignment are described in Methods in Enzymology (vol. 266: Computer Methods for Macromolecular Sequence Analysis (1996), ed. Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA). An alignment program that permits gaps in the sequence may be utilized to align the sequences. For example, the Smith-Waterman is one type of algorithm that permits gaps in sequence alignments (*see, e.g.,* Meth. Mol. Biol. 70:173-187, 1997). Also, the GAP program using the Needleman and Wunsch alignment method may be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves the ability to detect distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Nucleic acid-encoded amino acid sequences may be used to search both protein and DNA databases. Databases with individual sequences are described in Methods in Enzymology, ed. Doolittle, *supra*. Databases include, for example, Genbank, EMBL, and DNA Database of Japan (DDBJ).

**[077]** As used herein, a nucleic acid or other molecule attached to an array is referred to as a "probe" or "capture probe." When an array contains several probes corresponding to one gene, these probes are referred to as a "gene-probe set." A gene-probe set may consist of, for example, about 2 to about 20 probes, from about 2 to about 10 probes, or about 5 probes.

**[078]** The "profile" of a cell's biological state refers to the levels of various constituents of a cell that are known to change in response to drug treatments and other perturbations of the biological state of the cell. Constituents of a cell include, for example, levels of RNA, levels of protein abundances, or protein activity levels.

**[079]** The term "protein," "polypeptide," and "peptide" are used interchangeably herein when referring to a gene product.

**[080]** An expression profile in one cell is "similar" to an expression profile in another cell when the level of expression of the genes in the two profiles are sufficiently similar that the similarity is indicative of a common characteristic, for example, the same type of cell. Accordingly, the expression profiles of a first cell and a second cell are similar when at least 75% of the genes that are expressed in the first cell are expressed in the second cell at a level that is within a factor of two relative to the first cell.

**[081]** "Small molecule," as used herein, refers to a composition with a molecular weight of less than about 5 kD. Small molecules can be nucleic acids, peptides, polypeptides, peptidomimetics, carbohydrates, lipids, or other organic or inorganic molecules. Many pharmaceutical companies have extensive libraries of chemical and/or biological mixtures, often fungal, bacterial, or algal extracts, which can be screened with any of the assays of the invention to identify compounds that modulate a bioactivity.

**[082]** The term "specific hybridization" of a probe to a target site of a template nucleic acid refers to hybridization of the probe predominantly to the target, such that the hybridization signal can be clearly interpreted. As further described herein, such conditions resulting in specific hybridization vary depending on the length of the region of homology, the GC content of the region, and the melting temperature ("Tm") of the hybrid. Thus, hybridization conditions may vary in salt content, acidity, and temperature of the hybridization solution and the washes.

**[083]** A "variant" of polypeptide refers to a polypeptide having an amino acid sequence in which one or more amino acid residues is altered. The variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties (e.g., replacement of leucine with isoleucine). A variant may also have "nonconservative" changes (e.g., replacement of glycine with tryptophan). Analogous minor variations may include amino acid deletions or insertions, or both. Guidance in determining which amino acid residues may be substituted, inserted, or deleted without abolishing biological or immunological activity may be identified using computer programs well known in the art, for example, LASERGENE software (DNASTAR).

**[084]** The term "variant," when used in the context of a polynucleotide sequence, may encompass a polynucleotide sequence related to that of a particular gene or the coding sequence thereof. This definition may also include, for example, "allelic," "splice," "species," or "polymorphic" variants. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing of exons during mRNA processing. The corresponding polypeptide may possess additional functional domains or an absence of domains. Species variants are polynucleotide sequences that vary from one species to another. The resulting polypeptides generally will have significant amino acid identity relative to each other. A polymorphic variant is a variation in the polynucleotide sequence of a particular gene between individuals of a given species. Polymorphic variants also may encompass "single nucleotide polymorphisms" (SNPs) in which the polynucleotide sequence varies by one base. The presence of SNPs may be indicative of, for example, a certain population, a disease state, or a propensity for a disease state.

*Microarrays for Determining the Level of Expression of Genes*

**[085]** Generally, determining expression profiles with microarrays involves the following steps: (a) obtaining an mRNA sample from a subject and preparing labeled nucleic acids therefrom (the "target nucleic acids" or "targets"); (b) contacting the target nucleic acids with an array under conditions sufficient for the target nucleic acids to bind to the corresponding probes on the array, for example, by hybridization or specific binding; (c) optional removal of unbound targets from the array; (d) detecting the bound targets, and (e) analyzing the results, for example, using computer based analysis methods. As used herein, "nucleic acid probes" or "probes" are nucleic acids attached to the array, whereas "target nucleic acids" are nucleic acids that are hybridized to the array. Each of these steps is described in more detail below.

**[086]** Nucleic acid specimens may be obtained from an individual to be tested using either "invasive" or "non-invasive" sampling means. A sampling means is said to be "invasive" if it involves the collection of nucleic acids from within the skin or organs of an animal (including murine, human, ovine, equine, bovine, porcine, canine, or feline animal). Examples of invasive methods include blood collection, semen collection, needle biopsy, pleural aspiration, umbilical cord biopsy, etc. Examples of such methods are discussed by Kim, et al., (J. Virol. 66:3879-3882, 1992); Biswas, et al., (Ann. NY Acad. Sci. 590:582-583, 1990); and Biswas, et al., (J. Clin. Microbiol. 29:2228-2233, 1991).

**[087]** In contrast, a "non-invasive" sampling means is one in which the nucleic acid molecules are recovered from an internal or external surface of the animal. Examples of such "non-invasive" sampling means include, for example, "swabbing," collection of tears, saliva, urine, fecal material, sweat or perspiration, hair, etc.

**[088]** In one embodiment of the present invention, one or more cells from the subject to be tested are obtained and RNA is isolated from the cells. In one embodiment, a sample of peripheral blood leukocytes (PBLs) cells is obtained from the subject. It is also possible to obtain a cell sample from a subject, and then to enrich the sample for a desired cell type. For example, cells may be isolated from other cells using a variety of techniques, such as isolation with an antibody binding to an epitope on the cell surface of the desired cell type. Where the desired cells are in a solid tissue, particular cells may be dissected, for example, by microdissection or by laser capture microdissection (LCM) (*see, e.g.*, Bonner, et al., Science 278:1481, 1997; Emmert-Buck, et al., Science 274:998, 1996; Fend, et al., Am. J. Path. 154:61, 1999; and Murakami, et al., Kidney Int. 58:1346, 2000).

**[089]** RNA may be extracted from tissue or cell samples by a variety of methods, for example, guanidium thiocyanate lysis followed by CsCl centrifugation (Chirgwin, et al., Biochemistry

18:5294-5299, 1979). RNA from single cells may be obtained as described in methods for preparing cDNA libraries from single cells (*see, e.g.,* Dulac, Curr. Top. Dev. Biol. 36:245, 1998; Jena, et al., J. Immunol. Methods 190:199, 1996).

**[090]** The RNA sample can be further enriched for a particular species. In one embodiment, for example, poly(A)+ RNA may be isolated from an RNA sample. In particular, poly-T oligonucleotides may be immobilized on a solid support to serve as affinity ligands for mRNA. Kits for this purpose are commercially available, for example, the MessageMaker kit (Life Technologies, Grand Island, NY).

**[091]** In one embodiment, the RNA population may be enriched for sequences of interest, such as the genes characteristic of human breast tumor tissue (e.g., SEQ ID NOs: 1-127). Enrichment may be accomplished, for example, by primer-specific cDNA synthesis, or multiple rounds of linear amplification based on cDNA synthesis and template-directed *in vitro* transcription (*see, è.g.*, Wang, et al., Proc. Natl. Acad. Sci. USA 86:9717, 1989; Dulac, et al., *supra*; Jena, et al., *supra*).

**[092]** The population of RNA, enriched or not in particular species or sequences, may be further amplified. Such amplification is particularly important when using RNA from a single cell or a few cells. A variety of amplification methods are suitable for use in the methods of the present invention, including, for example, PCR; ligase chain reaction (LCR) (*see, e.g.*, Wu and Wallace, Genomics 4:560, 1989; Landegren, et al., Science 241:1077, 1988); self-sustained sequence replication (SSR) (*see, e.g.*, Guatelli, et al., Proc. Natl. Acad. Sci. USA 87:1874, 1990); nucleic acid based sequence amplification (NASBA) and transcription amplification (*see, e.g.*, Kwoh, et al., Proc. Natl. Acad. Sci. USA 86:1173, 1989). Methods for PCR technology are well known in the art (*see, e.g.*, PCR Technology: Principles and Applications for DNA Amplification (ed. H. A. Erlich, Freeman Press, N.Y., N.Y., 1992); PCR Protocols: A Guide to Methods and Applications (eds. Innis, et al., Academic Press, San Diego, Calif., 1990); Mattila, et al., Nucleic Acids Res. 19:4967, 1991; Eckert, et al., PCR Methods and Applications 1:17, 1991; PCR (eds. McPherson, et al., IRL Press, Oxford); and U.S. Pat. No. 4,683,202). Methods of amplification are described, for example, by Ohyama, et al., (BioTechniques 29:530, 2000); Luo, et al., (Nat. Med. 5:117, 1999); Hegde, et al., (BioTechniques 29:548, 2000); Kacharmina, et al., (Meth. Enzymol. 303:3, 1999); Livesey, et al., Curr. Biol. 10:301, 2000); Spirin, et al., (Invest. Ophtalmol. Vis. Sci. 40:3108, 1999); and Sakai, et al., (Anal. Biochem. 287:32, 2000). RNA amplification and cDNA synthesis may also be conducted in cells *in situ* (*see, e.g.*, Eberwine, et al. Proc. Natl. Acad. Sci. USA 89:3010, 1992).

**[093]** The target molecules may be labeled to permit detection of hybridization of the target molecules to a microarray. That is, the probe may comprise a member of a signal producing system and thus, is detectable, either directly or through combined action with one or more

additional members of a signal producing system. Examples of directly detectable labels include isotopic and fluorescent moieties incorporated, usually by a covalent bond, into a moiety of the probe, such as a nucleotide monomeric unit (e.g., dNMP of the primer), or a photoactive or chemically active derivative of a detectable label which can be bound to a functional moiety of the probe molecule.

**[094]** Nucleic acids may be labeled during or after enrichment and/or amplification of RNAs. For example, reverse transcription may be carried out in the presence of a dNTP conjugated to a . detectable label, for example, a fluorescently labeled dNTP. In another embodiment, the cDNA or RNA probe may be synthesized in the absence of detectable label and may be labeled subsequently, for example, by incorporating biotinylated dNTPs or rNTP, or some similar means (e.g., photo-cross-linking a psoralen derivative of biotin to RNAs), followed by addition of labeled streptavidin (e.g., phycoerythrin-conjugated streptavidin) or the equivalent.

**[095]** Fluorescent moieties or labels of interest include coumarin and its derivatives (e.g., 7-amino-4-methylcoumarin, aminocoumarin); bodipy dyes such as Bodipy FL and cascade blue; fluorescein and its derivatives (e.g., fluorescein isothiocyanate, Oregon green); rhodamine dyes (e.g., Texas red, tetramethylrhodamine); eosins and erythrosins; cyanine dyes (e.g., Cy2, Cy3, Cy3.5, Cy5, Cy5.5, Cy7); FluorX, macrocyclic chelates of lanthanide ions (e.g., quantum dye$^{TM}$); fluorescent energy transfer dyes such as thiazole orange-ethidium heterodimer, TOTAB, dansyl, etc. Individual fluorescent compounds which have functionalities for linking to an element desirably detected in an apparatus or assay of the invention, or which may be modified to incorporate such functionalities may also be utilized (*see, e.g.*, Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, Calif.).

**[096]** Chemiluminescent labels include luciferin and 2,3-dihydrophthalazinediones, for example, luminol.

**[097]** Labels may also be members of a signal producing system that act in concert with one or more additional members of the same system to provide a detectable signal. Illustrative of such labels are members of a specific binding pair, such as ligands, for example, biotin, fluorescein, digoxigenin, antigen, polyvalent cations, chelator groups and the like. Members may specifically bind to additional members of the signal producing system, and the additional members may provide a detectable signal either directly or indirectly, for example, an antibody conjugated to a fluorescent moiety or an enzymatic moiety capable of converting a substrate to a chromogenic product (e.g., alkaline phosphatase conjugate antibody and the like).

**[098]** Additional labels of interest include those that provide a signal only when the probe with which it is associated is specifically bound to a target molecule. Such labels include "molecular

beacons" as described in Tyagi and Kramer (Nature Biotech. 14:303, 1996) and EP 0 070 685 B1. Other labels of interest include those described in U.S. Patent No. 5,563,037; WO 97/17471; and WO 97/17076.

**[099]** In other embodiments, the target nucleic acid may not be labeled. In this case, hybridization may be determined, for example, by plasmon resonance (*see, e.g.*, Thiel, et al., Anal. Chem. 69:4948, 1997).

**[100]** In one embodiment, a plurality (e.g., 2, 3, 4, 5, or more) of sets of target nucleic acids are labeled and used in one hybridization reaction ("multiplex" analysis). For example, one set of nucleic acids may correspond to RNA from one cell and another set of nucleic acids may correspond to RNA from another cell. The plurality of sets of nucleic acids may be labeled with different labels, for example, different fluorescent labels (e.g., fluorescein and rhodamine) which have distinct emission spectra so that they can be distinguished. The sets may then be mixed and hybridized simultaneously to one microarray (*see, e.g.,* Shena, et al., Science 270:467-470, 1995).

**[101]** Examples of distinguishable labels for use when hybridizing a plurality of target nucleic acids to one array are well known in the art and include: two or more different emission wavelength fluorescent dyes such as Cy3 and Cy5; combination of fluorescent proteins and dyes such as phicoerythrin and Cy5; two or more isotopes with different energy of emission such as $^{32}P$ and $^{33}P$; gold or silver particles with different scattering spectra; labels which generate signals under different treatment conditions such as temperature, pH, treatment with additional chemical agents, etc.; or generate signals at different time points after treatment. Using one or more enzymes for signal generation allows for the use of an even greater variety of distinguishable labels, based on different substrate specificity of enzymes (e.g., alkaline phosphatase/peroxidase).

**[102]** The quality of labeled nucleic acids may be evaluated prior to hybridization to an array. In one embodiment, the GeneChip® Test3 Array from Affymetrix (Santa Clara, CA) may be used for that purpose. This array contains probes representing a subset of characterized genes from several organisms including mammals. Thus, the quality of a labeled nucleic acid sample can be determined by hybridization of a fraction of the sample to an array.

**[103]** Microarrays for use according to the invention include one or more probes of genes characteristic of human breast tumor tissue. In one embodiment, the microarray comprises one or more probes corresponding to one or more of genes selected from the group consisting of genes which are up-regulated in cancer and genes which are down-regulated in cancer. The microarray may comprise probes corresponding to, for example, at least 10, at least 20, at least 50, at least 100, or at least 1000 genes characteristic of human breast tumor tissue. The microarray may comprise probes corresponding to each gene listed in the Tables.

**[104]** There may be one or more than one probe corresponding to each gene on a microarray. For example, a microarray may contain from 2 to 20 probes corresponding to one gene or about 5 to 10. The probes may correspond to the full-length RNA sequence or complement thereof of genes characteristic of human breast tumor tissue, or the probe may correspond to a portion thereof, which portion is of sufficient length to permit specific hybridization. Such probes may comprise from about 50 nucleotides to about 100, 200, 500, or 1000 nucleotides or more than 1000 nucleotides. As further described herein, microarrays may contain oligonucleotide probes, consisting of about 10 to 50 nucleotides, about 15 to 30 nucleotides, or about 20-25 nucleotides. The probes may be single-stranded and will have sufficient complementarity to its target to provide for the desired level of sequence specific hybridization.

**[105]** Typically, the arrays used in the present invention will have a site density of greater than 100 different probes per $cm^2$. The arrays may have a site density of, for example, greater than 500/$cm^2$, greater than about 1000/$cm^2$, or greater than about 10,000/$cm^2$. The arrays may have, for example, more than about 100 different probes on a single substrate, more than about 1000 different probes, more than about 10,000 different probes, or more than about 100,000 different probes on a single substrate.

**[106]** A number of different microarray configurations and methods for their production are known to those of skill in the art and are disclosed in U.S. Patent Nos: 5,242,974; 5,384,261; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,445,934; 5,556,752; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,472,672; 5,527,681; 5,529,756; 5,545,531; 5,554,501; 5,561,071; 5,571,639; 5,593,839; 5,624,711; 5,700,637; 5,744,305; 5,770,456; 5,770,722; 5,837,832; 5,856,101; 5,874,219; 5,885,837; 5,919,523; 6,022,963; 6,077,674; and 6,156,501; Shena, et al., Tibtech 16:301, 1998; Duggan, et al., Nat. Genet. 21:10, 1999; Bowtell, et al., Nat. Genet. 21:25, 1999; Lipshutz, et al., 21 Nature Genet. 20-24, 1999; Blanchard, et al., 11 Biosensors and Bioelectronics, 687-90, 1996; Maskos, et al., 21 Nucleic Acids Res. 4663-69, 1993; Hughes, et al., Nat. Biotechol. 19:342, 2001; the disclosures of which are herein incorporated by reference. Patents describing methods of using arrays in various applications include: U.S. Pat. Nos. 5,143,854; 5,288,644; 5,324,633; 5,432,049; 5,470,710; 5,492,806; 5,503,980; 5,510,270; 5,525,464; 5,547,839; 5,580,732; 5,661,028; 5,848,659; and 5,874,219; the disclosures of which are herein incorporated by reference.

**[107]** Arrays may include control and reference nucleic acids. Control nucleic acids include, for example, prokaryotic genes such as bioB, bioC and bioD, cre from P1 bacteriophage or polyA controls, such as dap, lys, phe, thr, and trp. Reference nucleic acids allow the normalization of results from one experiment to another and the comparison of multiple experiments on a

quantitative level. Exemplary reference nucleic acids include housekeeping genes of known expression levels, for example, GAPDH, hexokinase, and actin.

**[108]** In one embodiment, an array of oligonucleotides may be synthesized on a solid support. Exemplary solid supports include paper, membranes, filters, pins, glass, plastics, polymers, metals, metalloids, ceramics, organics, etc. Using chip masking technologies and photoprotective chemistry, it is possible to generate ordered arrays of nucleic acid probes. These arrays, which are known, for example, as "DNA chips" or very large scale immobilized polymer arrays ("VLSIPS$^{TM}$" arrays), may include millions of defined probe regions on a substrate having an area of about 1 $cm^2$ to several $cm^2$, thereby incorporating from a few to millions of probes (*see, e.g.*, U.S. Patent No. 5,631,734).

**[109]** A nucleic acid probe may be at least, for example, about 10, 15, 20, 25, 30, 50, 100 or more nucleotides, and may comprise the full-length gene. For example, probes may be those that hybridize specifically to the genes listed in the Tables.

**[110]** Nucleic acid probes may be obtained, for example, by PCR amplificartion of gene segments from genomic, cDNA (e.g., RT-PCR), or cloned sequences. cDNA probes may be prepared according to methods known in the art and further described herein, for example, by reverse-transcription PCR (RT-PCR) of RNA using sequence specific primers. Sequences of genes or cDNA from which probes are generated may be obtained, for example, from GenBank, other public databases, or publications.

**[111]** Oligonucleotide probes may also be synthesized by standard methods known in the art, for example, by automated DNA synthesizer or any other chemical method. As an example, phosphorothioate oligonucleotides may be synthesized by the method of Stein, et al., (Nucl. Acids Res. 16:3209, 1988), and methylphosphonate oligonucleotides may be prepared by controlled pore glass polymer supports (*see, e.g.,* Sarin, et al., Proc. Natl. Acad. Sci. U.S.A. 85:7448-7451, 1988). In another embodiment, the oligonucleotide may be a 2'-)-methylribonucleotide (Inoue, et al., Nucl. Acids Res. 15:6131-6148, 1987), or a chimeric RNA-DNA analog (Inoue, et al., FEBS Lett. 215:327-330, 1987).

**[112]** Nucleic acid probes may be natural nucleic acids or chemically modified nucleic acids (e.g., composed of nucleotide analogs); however, the probes should possess activated hydroxyl groups compatible with the linking chemistry. The protective groups may be photolabile, or the protective groups may be labile under certain chemical conditions (e.g., acid). The surface of the solid support may contain a composition that generates acids upon exposure to light. Thus, exposure of a region of the substrate to light generates acids in that region that remove the protective groups in the exposed region. Also, the synthesis method may use 3'- protected 5'-0-phosphoramidite-

activated deoxynucleoside. In this case, the oligonucleotide is synthesized in the 5' to 3' direction, which results in a free 5' end.

**[113]** In one embodiment of the present invention, oligonucleotides of an array may be synthesized using a 96-well automated multiplex oligonucleotide synthesizer (A.M.O.S.) that is capable of producing thousands of oligonucleotides (*see, e.g.,* Lashkari, et al., Proc. Natl. Acad. Sci. USA 93: 7912, 1995).

**[114]** To compare expression levels, labeled nucleic acids may be contacted with the array under conditions sufficient for binding between the target nucleic acid and the probe on the array. In one embodiment, the hybridization conditions may be selected to provide for the desired level of hybridization specificity; that is, conditions sufficient for hybridization to occur between the labeled nucleic acids and probes on the microarray.

**[115]** Hybridization may be carried out in conditions permitting essentially specific hybridization. The length and GC content of the nucleic acid will determine the thermal melting point and thus, the hybridization conditions necessary for obtaining specific hybridization of the probe to the target nucleic acid. These factors are well known to a person of skill in the art, and may also be tested in assays. An extensive guide to nucleic acid hybridization may be found in Tijssen, et al. (Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)). Generally, stringent conditions may be selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Highly stringent conditions may be selected to be equal to the Tm point for a particular probe. Sometimes the term "dissociation temperature" (Td) is used to define the temperature at which at least half of the probe dissociates from a perfectly matched target nucleic acid. In any case, a variety of techniques for estimating the Tm or Td are available, and generally are described in Tijssen, *supra*. Typically, G-C base pairs in a duplex are estimated to contribute about 3°C to the Tm, while A-T base pairs are estimated to contribute about 2°C, up to a theoretical maximum of about 80-100°C. However, more sophisticated models of Tm and Td are available in which G-C stacking interactions, solvent effects, the desired assay temperature, and the like are taken into account.

**[116]** In one embodiment, non-specific binding or background signal may be reduced by the use of a detergent (e.g, C-TAB) or a blocking reagent (e.g., sperm DNA, cot-1 DNA, etc.) during the hybridization. In another embodiment, the hybridization may be performed in the presence of about 0.5 mg/ml DNA (e.g., herring sperm DNA). The use of blocking agents in hybridization is well known to those of skill in the art (*see, e.g.,* Tijssen, *supra*).

**[117]** If the target sequences are detected using the same label, different arrays may be employed for each physiological source or the same array may be screened multiple times. The above methods may be varied to provide for multiplex analysis by employing different and distinguishable labels for the different target populations (e.g., different physiological sources). According to this multiplex method, the same array may be used at the same time for each of the different target populations.

**[118]** The methods described above result in the production of hybridization patterns of labeled target nucleic acids on the array surface. The resultant hybridization patterns of labeled nucleic acids may be visualized or detected in a variety of ways, with the particular manner of detection selected based on the particular label of the target nucleic acid. Representative detection means include scintillation counting, autoradiography, fluorescence measurement, colorimetric measurement, light emission measurement, light scattering, and the like.

**[119]** One such method of detection utilizes an array scanner that is commercially available (Affymetrix, Santa Clara, CA), for example, the 417™ Arrayer, the 418™ Array Scanner, or the Agilent GeneArray™ Scanner. This scanner is controlled from a system computer with an interface and easy-to-use software tools. The output may be directly imported into or directly read by a variety of software applications. Scanning devices are described in, for example, U.S. Patent Nos. 5,143,854 and 5,424,186.

**[120]** For fluorescent labeled probes, the fluorescence emissions at each site of a transcript array may be, for example, detected by scanning confocal laser microscopy. Alternatively, a laser may be used that allows simultaneous specimen illumination at wavelengths specific to the two fluorophores and emissions from the two fluorophores may be analyzed simultaneously (*see, e.g.,* Shalon, et al., Genome Res. 6:639-645, 1996). In one embodiment, the arrays may be scanned with a laser fluorescent scanner with a computer controlled X-Y stage and a microscope objective. Fluorescence laser scanning devices are described in Shalon, et al., *supra*.

**[121]** Following the data gathering operation, the data will typically be reported to a data analysis operation. To facilitate the sample analysis operation, the data obtained by the reader from the device may be analyzed using a digital computer. Typically, the computer will be appropriately programmed for receipt and storage of the data from the device, as well as for analysis and reporting of the data gathered, for example, subtraction of the background, deconvolution of multi-color images, flagging or removing artifacts, verifying that controls have performed properly, normalizing the signals, interpreting fluorescence data to determine the amount of hybridized target, normalization of background and single base mismatch hybridizations, and the like.

**[122]** In one embodiment, a system comprises a search function that allows one to search for specific patterns, for example, patterns relating to differential gene expression, for example, between the expression profile of a cancer cell and the expression profile of a counterpart normal cell in a subject. A system may also allow one to search for patterns of gene expression between more than two samples.

**[123]** Various algorithms are available for analyzing gene expression profile data, for example, the type of comparisons to perform. In certain embodiments, it is desirable to group genes that are co-regulated. This allows for the comparison of large numbers of profiles. One embodiment for identifying such groups of genes involves clustering algorithms (for reviews of clustering algorithms, *see, e.g.*, Fukunaga, 1990, Statistical Pattern Recognition, 2nd Ed., Academic Press, San Diego; Everitt, 1974, Cluster Analysis, London: Heinemann Educ. Books; Hartigan, 1975, Clustering Algorithms, New York: Wiley; Sneath and Sokal, 1973, Numerical Taxonomy, Freeman; Anderberg, 1973, Cluster Analysis for Applications, Academic Press: New York).

**[124]** Clustering may be based on other characteristics of the genes, for example, their level of expression (*see, e.g.*, U.S. Patent No. 6,203,987), or permit clustering of time curves (*see, e.g.* U.S. Patent No. 6,263,287). Examples of clustering algorithms include K-means clustering and hierarchical clustering. Clustering may also be achieved by visual inspection of gene expression data using a graphical representation of the data (e.g. a "heat map"). An example of software which contains clustering algorithms and a means to graphically represent gene expression data is Spotfire DecisionSite (Spotfire, Inc., Somerville, Massachusetts and Göteborg, Sweden).

**[125]** Comparison of the expression levels of one or more genes characteristic of human breast tumor tissue with reference expression levels, for example, expression levels in diseased cells of cancer or in normal counterpart cells, may be conducted using computer systems. In one embodiment, expression levels may be obtained from two cells and these two sets of expression levels may be introduced into a computer system for comparison. In another embodiment, one set of expression levels is entered into a computer system for comparison with values that are already present in the computer system, or in computer-readable form that is then entered into the computer system.

**[126]** In one embodiment, the computer system may also contain a database comprising values representing levels of expression of one or more genes characteristic of human breast tumor tissue. The database may contain one or more expression profiles of genes characteristic of human breast tumor tissue in different cells.

**[127]** In another embodiment, the invention provides a computer-readable form of the gene expression profile data, or of values corresponding to the level of expression of at least one gene

characteristic of cancer in a diseased cell. The values may be mRNA expression levels obtained from experiments, for example, microarray analysis. The values may also be mRNA levels normalized relative to a reference gene whose expression is constant in numerous cells under numerous conditions (e.g., GAPDH). In other embodiments, the values in the computer may be ratios of, or differences between, normalized or non-normalized mRNA levels in different samples.

**[128]** In one embodiment, the expression data of a cell of a subject treated *in vitro* or *in vivo* with the drug is entered into a computer and the computer is instructed to compare the data entered to the data in the computer, and to provide results indicating whether the expression data input into the computer are more similar to those of a cell of a subject that is responsive to the drug or more similar to those of a cell of a subject that is not responsive to the drug. Thus, the results indicate whether the subject is likely to respond to the treatment with the drug or unlikely to respond to it.

**[129]** The invention also provides a machine-readable or computer-readable medium including program instructions for performing the following steps: (i) comparing a plurality of values corresponding to expression levels of one or more genes characteristic of human breast tumor tissue in a query cell with a database including records comprising reference expression or expression profile data of one or more reference cells and an annotation of the type of cell; and (ii) indicating to which cell the query cell is most similar based on similarities of expression profiles. The reference cells may be cells from subjects at different stages of cancer. The reference cells may also be cells from subjects responding or not responding to a particular drug treatment and optionally incubated *in vitro* or *in vivo* with the drug.

**[130]** The reference cells may also be cells from subjects responding or not responding to several different treatments, and the computer system indicates a preferred treatment for the subject. Accordingly, the invention provides a method for selecting a therapy for a patient having cancer, the method comprising: (i) providing the level of expression of one or more genes characteristic of human breast tumor tissue in a diseased cell of the patient; (ii) providing a plurality of reference profiles, each associated with a therapy, wherein the subject expression profile and each reference profile has a plurality of values, each value representing the level of expression of a gene characteristic of cancer; and (iii) selecting the reference profile most similar to the subject expression profile, to thereby select a therapy for said patient. In one embodiment, step (iii) is performed by a computer. The most similar reference profile may be selected by weighing a comparison value of the plurality using a weight value associated with the corresponding expression data.

**[131]** The relative abundance of an mRNA in two biological samples may be scored as a perturbation and its magnitude determined (i.e., the abundance is different in the two sources of

mRNA tested), or as not perturbed (i.e., the relative abundance is the same). In various embodiments, a difference between the two sources of RNA of at least a factor of about 25% (RNA from one source is 25% more abundant in one source than the other source), more usually about 50%, even more often by a factor of about 2 (twice as abundant), 3 (three times as abundant) or 5 (five times as abundant) is scored as a perturbation. Perturbations may be used by a computer for calculating and expression comparisons.

*Drug Design Using Microarrays*

**[132]** The invention also provides methods for designing and optimizing drugs for cancer. In one embodiment, compounds may be screened by comparing the expression level of one or more genes characteristic of human breast tumor tissue following incubation of a diseased cell of cancer or similar cell with the test compound. In another embodiment, the expression level of the genes may be determined using microarrays, and comparing the gene expression profile of a cell in response to the test compound with the gene expression profile of a normal cell corresponding to a diseased cell of cancer (a "reference profile"). In a further embodiment, the expression profile may also be compared to that of a diseased cell of cancer. The comparisons may be done by introducing the gene expression profile data of the cell treated with drug into a computer system comprising reference gene expression profiles, which are stored in a computer readable form, using appropriate algorithms. Test compounds may be screened for those that alter the level of expression of genes characteristic of human breast tumor tissue. Such compounds, that is, compounds which are capable of normalizing the expression of essentially all genes characteristic of human breast tumor tissue, are candidate therapeutics.

**[133]** The efficacy of the compounds may then be tested in additional *in vitro* and *in vivo* assays, and in animal models (e.g., xenograft model). The test compound may be administered to the test animal, and one or more symptoms of the disease may be monitored for improvement of the condition of the animal. Expression of one or more genes characteristic of human breast tumor tissue may also be measured before and after administration of the test compound to the animal. A normalization of the expression of one or more of these genes is indicative of the efficiency of the compound for treating cancer in the animal.

**[134]** In the clinical setting, obtaining human-derived samples of tissue exhibiting cancer may be difficult, if not prohibitive. Therefore, identification of gene expression changes indicative of efficacy of a therapeutic compound may be determined in a more easily accessible, surrogate cell population, for example, peripheral blood leukocytes (PBLs). This method may be performed either in a human or animal model system. In one embodiment, a test compound may be administered to the test animal (either normal or cancer-containing) at the same doses that have been observed to be efficacious in treating cancer in that animal model. Blood may be drawn from

the animal at various time points (e.g., 1, 4, 7, and 24 hours following the first, mid-point, and last day of a regimen of multiple day dosing). Animals dosed with vehicle may be used as controls. RNA may be isolated from PBLs, and can be used to generate probes for hybridization to microarrays. The hybridization results may then be analyzed using computer programs and databases, as described above. The resulting expression profile may be compared directly to the analogous profile from the treated cancer tissue for similarities or simply correlated with efficacy (e.g., in terms of doses and time points) in the animal model.

**[135]** In another embodiment, human blood may be treated *ex vivo* with a therapeutic compound at a dose consistent with the therapeutic dose in the animal model, or at a dose that is consistent with known plasma levels of the therapeutic dose in the animal model. The blood may be treated (e.g., rocking at 37°C) with the therapeutic compound immediately, or after some period of incubation time (e.g., 24 hours) to allow for gene expression to re-equilibrate after the blood draw. The blood may also be treated with the therapeutic compound for various timepoints (e.g., 4 and 24 hours), and then PBL RNA isolated and used to create a probe for hybridization to a microarray. A compound solubilization agent (e.g., DMSO) may be used as a control. The resulting expression profile may be compared directly to the analogous profile from the treated cancer tissue for similarities or simply correlated with efficacy (e.g., in terms of doses and time points) in the animal model.

**[136]** The toxicity of the candidate therapeutic compound may be evaluated, for example, by determining whether the compound induces the expression of genes known to be associated with a toxic response. Expression of such toxicity related genes may be determined in different cell types, for example, those that are known to express the genes. In fact, alterations in gene expression may serve as a more sensitive marker of human toxicity than routine preclinical safety studies. In one method, microarrays may be used for detecting changes in the expression of genes known to be associated with a toxic response. It may be possible to perform proof of concept studies demonstrating that changes in gene expression levels may predict toxic events that were not identified by routine preclinical safety testing (*see, e.g.*, Huang, et al., Toxicol. Sci. 63:196-207, 2001; Waring, et al., Toxicol. Appl. Pharamacol. 175:28-42, 2001).

**[137]** Drug screening may be performed by adding a test compound to a sample of cells, and monitoring the effect. A parallel sample which does not receive the test compound may also be monitored as a control. The treated and untreated cells are then compared by any suitable phenotypic criteria, including but not limited to microscopic analysis, viability testing, ability to replicate, histological examination, the level of a particular RNA or polypeptide associated with the cells, the level of enzymatic activity expressed by the cells or cell lysates, and the ability of the

cells to interact with other cells or compounds. Differences between treated and untreated cells indicates effects attributable to the test compound.

**[138]** Desirable effects of a test compound include an effect on any phenotype that was conferred by the cancer-associated marker nucleic acid sequence. Examples include a test compound that limits the overabundance of mRNA, limits production of the encoded protein, or limits the functional effect of the protein. The effect of the test compound would be apparent when comparing results between treated and untreated cells.

*Diagnostic and Prognostic Assays*

**[139]** The present invention provides nucleic acid sequences which are differentially regulated in cancer, and a method for identifying such sequences. The present invention provides a method for identifying a nucleotide sequence which is differentially regulated in a subject with cancer, comprising: hybridizing a nucleic acid sample corresponding to RNA obtained from the subject to a nucleic acid sample comprising one or more nucleic acid molecules of known identity; and measuring the hybridization of the nucleic acid sample to the one or more nucleic acid molecules of known identity, wherein a difference in the hybridization of the nucleic acid sample to the one or more nucleic acid molecules of known identity relative to a nucleic acid sample obtained from a subject without cancer is indicative of the differential expression of the nucleotide sequence in a subject with cancer.

**[140]** Generally, the present invention provides a method for identifying nucleic acid sequences which are differentially regulated in a subject with cancer comprising isolating messenger RNA from a subject, generating cRNA from the mRNA sample, hybridizing the cRNA to a microarray comprising a plurality of nucleic acid molecules stably associated with discrete locations on the array, and identifying patterns of hybridization of the cRNA to the array. According to the present invention, a nucleic acid molecule which hybridizes to a given location on the array is said to be differentially regulated if the hybridization signal is, for example, at least two-fold higher or lower than the hybridization signal at the same location on an identical array hybridized with a nucleic acid sample obtained from a subject that does not have cancer.

**[141]** Expression patterns may be used to derive a panel of biomarkers that can be used to predict the efficacy of drug treatment in the patients. The biomarkers may consist of gene expression levels from microarray experiments on RNA isolated from biological samples, RNA isolated from frozen samples of tumor biopsies, or mass spectrometry-derived protein masses in the serum.

**[142]** Although the precise mechanism for data analysis will depend upon the exact nature of the data, a typical procedure for developing a panel of biomarkers is as follows. The data (gene expression levels or mass spectra) are collected for each patient prior to treatment. As the study

progresses, the patients are classified according to their response to the drug treatment; either as efficacious or non-efficacious. Multiple levels of efficacy can be accommodated in a data model, but a binary comparison is considered optimal, particularly if the patient population is less than several hundred. Assuming adequate numbers of patients in each class, the protein and/or gene expression data may be analyzed by a number of techniques known in the art. Many of the techniques are derived from traditional statistics as well from the field of machine learning. These techniques serve two purposes:

1. Reduce the dimensionality of data - In the case of mass spectra or gene expression microarrays, data is reduced from many thousands of individual data points to bout three to ten. The reduction is based upon the predictive power of the data points when taken as a set.

2. Training - These three to ten data points are then used to train multiple machine learning algorithms which then "learn" to recognize, in this case, patterns of protein masses or gene expression which distinguish efficacious drug treatment from non-efficacious. All patient samples can be used to train the algorithms.

**[143]** The resulting trained algorithms are then tested in order to measure their predictive power. Typically, when less than many hundreds of training examples are available, some form of cross-validation is performed. To illustrate, consider a ten-fold cross validation. In this case, patient samples are randomly assigned to one of ten bins. In the first round of validation the samples in nine of the bins are used for training and the remaining samples in the tenth bin are used to test the algorithm. This is repeated an additional nine times, each time leaving out the samples in a different bin for testing. The results (correct predictions and errors) from all ten rounds are combined and the predictive power is then assessed. Different algorithms, as well as different panels, may be compared in this way for this study. The "best" algorithm/panel combination will then be selected. This "smart" algorithm may then be used in future studies to select the patients that are most likely to respond to treatment.

**[144]** Many algorithms benefit from additional information taken for the patients. For example, gender or age could be used to improve predictive power. Also, data transformations such as normalization and smoothing may be used to reduce noise. Because of this, a large number of algorithms may be trained using many different parameters in order to optimize the outcome. If predictive patterns exist in the data, it is likely that an optimal, or near-optimal, "smart" algorithm can be developed. If more patient samples become available, the algorithm can be retrained to take advantage of the new data.

**[145]** As an example using mass spectrometry, plasma may be applied to a hydrophobic SELDI-target, washed extensively in water, and analyzed by the SELDI-Tof mass spectrometer. This may be repeated on 100 or more patient samples. The protein profiles resulting from the intensities of

some 16,000 m/z values in each sample would be statistically analyzed in order to identify sets of specific m/z values that are predictive of drug efficacy. Identical experiments using other SELDI-targets, such as ion-exchange or IMAC surfaces, could also be conducted. These will capture different subsets of the proteins present in plasma. Furthermore, the plasma may be denatured and prefractionated prior to application onto the SELDI target.

**[146]** The present invention provides methods for determining whether a subject is at risk for developing a disease or condition characterized by unwanted cell proliferation by detecting biomarkers, that is, nucleic acids and/or polypeptide markers for cancer.

**[147]** In clinical applications, human tissue samples may be screened for the presence and/or absence of biomarkers identified herein. Such samples could consist of needle biopsy cores, surgical resection samples, lymph node tissue, or serum. For example, these methods include obtaining a biopsy, which is optionally fractionated by cryostat sectioning to enrich tumor cells to about 80% of the total cell population. In certain embodiments, nucleic acids extracted from these samples may be amplified using techniques well known in the art. The levels of selected markers detected would be compared with statistically valid groups of metastatic, non-metastatic malignant, benign, or normal tissue samples.

**[148]** In one embodiment, the diagnostic method comprises determining whether a subject has an abnormal mRNA and/or protein level of the biomarkers, such as by Northern blot analysis, reverse transcription-polymerase chain reaction (RT-PCR), *in situ* hybridization, immunoprecipitation, Western blot hybridization, or immunohistochemistry. According to the method, cells may be obtained from a subject and the levels of the biomarkers, protein, or mRNA level, are determined and compared to the level of these markers in a healthy subject. An abnormal level of the biomarker polypeptide or mRNA levels is likely to be indicative of cancer.

**[149]** In one embodiment, the method comprises using a nucleic acid probe to determine the presence of cancerous cells in a tissue from a patient. Specifically, the method comprises:

1. providing a nucleic acid probe comprising a nucleotide sequence, for example, at least 10, 15, 25 or 40 nucleotides, and up to all or nearly all of the coding sequence which is complementary to a portion of the coding sequence of a nucleic acid sequence and is differentially expressed in tumors cells;

2. obtaining a tissue sample from a patient potentially comprising cancerous cells;

3. providing a second tissue sample containing cells substantially all of which are non-cancerous;

4. contacting the nucleic acid probe under stringent conditions with RNA of each of said first and second tissue samples (e.g., in a Northern blot or *in situ* hybridization assay); and

5. comparing (a) the amount of hybridization of the probe with RNA of the first tissue sample, with (b) the amount of hybridization of the probe with RNA of the second tissue sample; wherein a statistically significant difference in the amount of hybridization with the RNA of the first tissue sample as compared to the amount of hybridization with the RNA of the second tissue sample is indicative of the presence of cancerous cells in the first tissue sample.

**[150]** In one aspect, the method comprises *in situ* hybridization with a probe derived from a given marker nucleic acid sequence. The method comprises contacting the labeled hybridization probe with a sample of a given type of tissue potentially containing cancerous or pre-cancerous cells as well as normal cells, and determining whether the probe labels some cells of the given tissue type to a degree significantly different (e.g., by at least a factor of two, or at least a factor of five, or at least a factor of twenty, or at least a factor of fifty) than the degree to which it labels other cells of the same tissue type.

**[151]** Also within the invention is a method of determining the phenotype of a test cell from a given human tissue, for example, whether the cell is (a) normal, or (b) cancerous or precancerous, by contacting the mRNA of a test cell with a nucleic acid probe, for example, at least about 10, 15, 25, or 40 nucleotides, and up to all or nearly all of a sequence which is complementary to a portion of the coding sequence of a nucleic acid sequence, and which is differentially expressed in tumor cells as compared to normal cells of the given tissue type; and determining the approximate amount of hybridization of the probe to the mRNA, an amount of hybridization either more or less than that seen with the mRNA of a normal cell of that tissue type being indicative that the test cell is cancerous or pre-cancerous.

**[152]** Alternatively, the above diagnostic assays may be carried out using antibodies to detect the protein product encoded by the marker nucleic acid sequence. Accordingly, in one embodiment, the assay would include contacting the proteins of the test cell with an antibody specific for the gene product of a nucleic acid, the marker nucleic acid being one which is expressed at a given control level in normal cells of the same tissue type as the test cell, and determining the approximate amount of immunocomplex formation by the antibody and the proteins of the test cell, wherein a statistically significant difference in the amount of the immunocomplex formed with the proteins of a test cell as compared to a normal cell of the same tissue type is an indication that the test cell is cancerous or pre-cancerous.

**[153]** The method for producing polyclonal and/or monoclonal antibodies which specifically bind to polypeptides useful in the present invention is known to those of skill in the art and may be found in, for example, Dymecki, et al., (J. Biol. Chem. 267:4815, 1992); Boersma & Van Leeuwen, (J. Neurosci. Methods 51:317, 1994); Green, et al., (Cell 28:477, 1982); and Arnheiter, et al., (Nature 294:278, 1981).

**[154]** Another such method includes the steps of: providing an antibody specific for the gene product of a marker nucleic acid sequence, the gene product being present in cancerous tissue of a given tissue type at a level more or less than the level of the gene product in non-cancerous tissue of the same tissue type; obtaining from a patient a first sample of tissue of the given tissue type, which sample potentially includes cancerous cells; providing a second sample of tissue of the same tissue type (which may be from the same patient or from a normal control, e.g. another individual or cultured cells), this second sample containing normal cells and essentially no cancerous cells; contacting the antibody with protein (which may be partially purified, in lysed but unfractionated cells, or *in situ*) of the first and second samples under conditions permitting immunocomplex formation between the antibody and the marker nucleic acid sequence product present in the samples; and comparing (a) the amount of immunocomplex formation in the first sample, with (b) the amount of immunocomplex formation in the second sample, wherein a statistically significant difference in the amount of immunocomplex formation in the first sample less as compared to the amount of immunocomplex formation in the second sample is indicative of the presence of cancerous cells in the first sample of tissue.

**[155]** The subject invention further provides a method of determining whether a cell sample obtained from a subject possesses an abnormal amount of marker polypeptide which comprises (a) obtaining a cell sample from the subject, (b) quantitatively determining the amount of the marker polypeptide in the sample so obtained, and (c) comparing the amount of the marker polypeptide so determined with a known standard, so as to thereby determine whether the cell sample obtained from the subject possesses an abnormal amount of the marker polypeptide. Such marker polypeptides may be detected by immunohistochemical assays, dot-blot assays, ELISA, and the like.

**[156]** Immunoassays are commonly used to quantitate the levels of proteins in cell samples, and many other immunoassay techniques are known in the art. The invention is not limited to a particular assay procedure, and therefore, is intended to include both homogeneous and heterogeneous procedures. Exemplary immunoassays which may be conducted according to the invention include fluorescence polarization immunoassay (FPIA), fluorescence immunoassay (FIA), enzyme immunoassay (EIA), nephelometric inhibition immunoassay (NIA), enzyme-linked immunosorbent assay (ELISA), and radioimmunoassay (RIA). An indicator moiety, or label

group, may be attached to the subject antibodies and is selected so as to meet the needs of various uses of the method which are often dictated by the availability of assay equipment and compatible immunoassay procedures. General techniques to be used in performing the various immunoassays noted above are known to those of ordinary skill in the art.

**[157]** In another embodiment, the level of the encoded product, or alternatively the level of the polypeptide, in a biological fluid (e.g., blood or urine) of a patient may be determined as a way of monitoring the level of expression of the marker nucleic acid sequence in cells of that patient. Such a method would include the steps of obtaining a sample of a biological fluid from the patient, contacting the sample (or proteins from the sample) with an antibody specific for an encoded marker polypeptide, and determining the amount of immune complex formation by the antibody, with the amount of immune complex formation being indicative of the level of the marker encoded product in the sample. This determination is particularly instructive when compared to the amount of immune complex formation by the same antibody in a control sample taken from a normal individual or in one or more samples previously or subsequently obtained from the same person.

**[158]** In another embodiment, the method may be used to determine the amount of marker polypeptide present in a cell, which in turn may be correlated with progression of a hyperproliferative disorder. The level of the marker polypeptide may be used predictively to evaluate whether a sample of cells contains cells which are, or are predisposed towards becoming, transformed cells. Moreover, the subject method may be used to assess the phenotype of cells which are known to be transformed, the phenotyping results being useful in planning a particular therapeutic regimen. For example, very high levels of the marker polypeptide in sample cells is a powerful diagnostic and prognostic marker for a cancer. The observation of marker polypeptide levels may be utilized in decisions regarding, for example, the use of more aggressive therapies.

**[159]** As set out above, one aspect of the present invention relates to diagnostic assays for determining, in the context of cells isolated from a patient, if the level of a marker polypeptide is significantly reduced in the sample cells. The term "significantly reduced" refers to a cell phenotype wherein the cell possesses a reduced cellular amount of the marker polypeptide relative to a normal cell of similar tissue origin. For example, a cell may have less than about 50%, 25%, 10%, or 5% of the marker polypeptide compared to that of a normal control cell. In particular, the assay evaluates the level of marker polypeptide in the test cells, and, preferably, compares the measured level with marker polypeptide detected in at least one control cell, for example, a normal cell and/or a transformed cell of known phenotype.

**[160]** Of particular importance to the subject invention is the ability to quantitate the level of marker polypeptide as determined by the number of cells associated with a normal or abnormal marker polypeptide level. The number of cells with a particular marker polypeptide phenotype

may then be correlated with patient prognosis. In one embodiment of the invention, the marker polypeptide phenotype of a lesion is determined as a percentage of cells in a biopsy which are found to have abnormally high/low levels of the marker polypeptide. Such expression may be detected by immunohistochemical assays, dot-blot assays, ELISA, and the like.

**[161]** Where tissue samples are employed, immunohistochemical staining may be used to determine the number of cells having the marker polypeptide phenotype. For such staining, a multiblock of tissue may be taken from the biopsy or other tissue sample and subjected to proteolytic hydrolysis, employing such agents as protease K or pepsin. In certain embodiments, it may be desirable to isolate a nuclear fraction from the sample cells and detect the level of the marker polypeptide in the nuclear fraction.

**[162]** The tissue samples are fixed by treatment with a reagent such as formalin, glutaraldehyde, methanol, or the like. The samples are then incubated with an antibody (e.g., a monoclonal antibody) with binding specificity for the marker polypeptides. This antibody may be conjugated to a label for subsequent detection of binding. Samples are incubated for a time sufficient for formation of the immunocomplexes. Binding of the antibody is then detected by virtue of a label conjugated to this antibody. Where the antibody is unlabeled, a second labeled antibody may be employed, for example, which is specific for the isotype of the anti-marker polypeptide antibody. Examples of labels which may be employed include radionuclides, fluorescers, chemiluminescers, enzymes, and the like.

**[163]** Where enzymes are employed, the substrate for the enzyme may be added to the samples to provide a colored or fluorescent product. Examples of suitable enzymes for use in conjugates include horseradish peroxidase, alkaline phosphatase, malate dehydrogenase, and the like. Where not commercially available, such antibody-enzyme conjugates are readily produced by techniques known to those skilled in the art.

**[164]** In one embodiment, the assay is performed as a dot blot assay. The dot blot assay finds particular application where tissue samples are employed as it allows determination of the average amount of the marker polypeptide associated with a single cell by correlating the amount of marker polypeptide in a cell-free extract produced from a predetermined number of cells.

**[165]** It is well established in the cancer literature that tumor cells of the same type (e.g., breast and/or colon tumor cells) may not show uniformly increased expression of individual oncogenes or uniformly decreased expression of individual tumor suppressor genes. There may also be varying levels of expression of a given marker gene even between cells of a given type of cancer, further emphasizing the need for reliance on a battery of tests rather than a single test. Accordingly, in

one aspect, the invention provides for a battery of tests utilizing a number of probes of the invention, in order to improve the reliability and/or accuracy of the diagnostic test.

**[166]** In one embodiment, the present invention also provides a method wherein nucleic acid probes are immobilized on a DNA chip in an organized array. Oligonucleotides may be bound to a solid support by a variety of processes, including lithography. For example, a chip may hold up to 250,000 oligonucleotides. These nucleic acid probes comprise a nucleotide sequence, for example, at least about 12, 15, 25, or 40 nucleotides in length, and up to all or nearly all of a sequence which is complementary to a portion of the coding sequence of a marker nucleic acid sequence and is differentially expressed in tumor cells. The present invention provides significant advantages over the available tests for various cancers, because it increases the reliability of the test by providing an array of nucleic acid markers on a single chip.

**[167]** The method includes obtaining a biopsy, which is optionally fractionated by cryostat sectioning to enrich tumor cells to about 80% of the total cell population. The DNA or RNA is then extracted, amplified, and analyzed with a DNA chip to determine the presence of absence of the marker nucleic acid sequences.

**[168]** In one embodiment, the nucleic acid probes are spotted onto a substrate in a two-dimensional matrix or array. Samples of nucleic acids may be labeled and then hybridized to the probes. Double-stranded nucleic acids, comprising the labeled sample nucleic acids bound to probe nucleic acids, may be detected once the unbound portion of the sample is washed away.

**[169]** The probe nucleic acids may be spotted on substrates including glass, nitrocellulose, etc. The probes can be bound to the substrate by either covalent bonds or by non-specific interactions, such as hydrophobic interactions. The sample nucleic acids can be labeled using radioactive labels, fluorophores, chromophores, etc.

**[170]** In yet another embodiment, the invention contemplates using a panel of antibodies which are generated against the marker polypeptides of this invention. Such a panel of antibodies may be used as a reliable diagnostic probe for cancer. The assay of the present invention comprises contacting a biopsy sample containing cells, for example, breast cells, with a panel of antibodies to one or more of the encoded products to determine the presence or absence of the marker polypeptides.

**[171]** The diagnostic methods of the subject invention may also be employed as follow-up to treatment, for example, quantitation of the level of marker polypeptides may be indicative of the effectiveness of current or previously employed cancer therapies as well as the effect of these therapies upon patient prognosis.

[172] In addition, the marker nucleic acids or marker polypeptides may be utilized as part of a diagnostic panel for initial detection, follow-up screening, detection of reoccurrence, and post-treatment monitoring for chemotherapy or surgical treatment.

[173] Accordingly, the present invention makes available diagnostic assays and reagents for detecting gain and/or loss of marker polypeptides from a cell in order to aid in the diagnosis and phenotyping of proliferative disorders arising from, for example, tumorigenic transformation of cells.

[174] The diagnostic assays described above may be adapted to be used as prognostic assays, as well. Such an application takes advantage of the sensitivity of the assays of the invention to events which take place at characteristic stages in the progression of a tumor. For example, a given marker gene may be up- or down-regulated at a very early stage, perhaps before the cell is irreversibly committed to developing into a malignancy, while another marker gene may be characteristically up- or down-regulated only at a much later stage. Such a method could involve the steps of contacting the mRNA of a test cell with a nucleic acid probe derived from a given marker nucleic acid which is expressed at different characteristic levels in cancerous or precancerous cells at different stages of tumor progression, and determining the approximate amount of hybridization of the probe to the mRNA of the cell, such amount being an indication of the level of expression of the gene in the cell, and thus an indication of the stage of tumor progression of the cell; alternatively, the assay may be carried out with an antibody specific for the gene product of the given marker nucleic acid, contacted with the proteins of the test cell. A battery of such tests will disclose not only the existence and location of a tumor, but also will allow the clinician to select the mode of treatment most appropriate for the tumor, and to predict the likelihood of success of that treatment.

[175] The methods of the invention may also be used to follow the clinical course of a tumor. For example, the assay of the invention may be applied to a tissue sample from a patient; following treatment of the patient for the cancer, another tissue sample is taken and the test repeated. Successful treatment will result in either removal of all cells which demonstrate differential expression characteristic of the cancerous or precancerous cells, or a substantial increase in expression of the gene in those cells, perhaps approaching or even surpassing normal levels.

[176] In yet another embodiment, the invention provides methods for determining whether a subject is at risk for developing a disease, such as a predisposition to develop cancer, associated with aberrant activity of a polypeptide, wherein the aberrant activity of the polypeptide is characterized by detecting the presence or absence of a genetic lesion characterized by at least one of (a) an alteration affecting the integrity of a gene encoding a marker polypeptides, or (b) the mis-expression of the encoding nucleic acid. To illustrate, such genetic lesions may be detected by

ascertaining the existence of at least one of (i) a deletion of one or more nucleotides from the nucleic acid sequence, (ii) an addition of one or more nucleotides to the nucleic acid sequence, (iii) a substitution of one or more nucleotides of the nucleic acid sequence, (iv) a gross chromosomal rearrangement of the nucleic acid sequence, (v) a gross alteration in the level of a messenger RNA transcript of the nucleic acid sequence, (vi) aberrant modification of the nucleic acid sequence, such as of the methylation pattern of the genomic DNA, (vii) the presence of a non-wild type splicing pattern of a messenger RNA transcript of the gene, (viii) a non-wild type level of the marker polypeptide, (ix) allelic loss of the gene, and/or (x) inappropriate post-translational modification of the marker polypeptide.

**[177]** The present invention provides assay techniques for detecting lesions in the encoding nucleic acid sequence. These methods include, but are not limited to, methods involving sequence analysis, Southern blot hybridization, restriction enzyme site mapping, and methods involving detection of absence of nucleotide pairing between the nucleic acid to be analyzed and a probe.

**[178]** Specific diseases or disorders, for example, genetic diseases or disorders, are associated with specific allelic variants of polymorphic regions of certain genes, which do not necessarily encode a mutated protein. Thus, the presence of a specific allelic variant of a polymorphic region of a gene in a subject may render the subject susceptible to developing a specific disease or disorder. Polymorphic regions in genes, may be identified, by determining the nucleotide sequence of genes in populations of individuals. If a polymorphic region is identified, then the link with a specific disease may be determined by studying specific populations of individuals, for example, individuals which developed a specific disease, such as cancer. A polymorphic region may be located in any region of a gene, for example, exons, in coding or non-coding regions of exons, introns, and promoter region.

**[179]** In an exemplary embodiment, there is provided a nucleic acid composition comprising a nucleic acid probe including a region of nucleotide sequence which is capable of hybridizing to a sense or antisense sequence of a gene or naturally occurring mutants thereof, or 5' or 3' flanking sequences or intronic sequences naturally associated with the subject genes or naturally occurring mutants thereof. The nucleic acid of a cell is rendered accessible for hybridization, the probe is contacted with the nucleic acid of the sample, and the hybridization of the probe to the sample nucleic acid is detected. Such techniques may be used to detect lesions or allelic variants at either the genomic or mRNA level, including deletions, substitutions, etc., as well as to determine mRNA transcript levels.

**[180]** An example of a detection method is allele specific hybridization using probes overlapping the mutation or polymorphic site and having about 5, 10, 20, 25, or 30 nucleotides around the mutation or polymorphic region. In one embodiment of the invention, several probes capable of

hybridizing specifically to allelic variants are attached to a solid phase support, for example, a "chip." Mutation detection analysis using these chips comprising oligonucleotides, also termed "DNA probe arrays" is described, for example, by Cronin, et al., (Human Mutation 7:244, 1996). In one embodiment, a chip may comprise all the allelic variants of at least one polymorphic region of a gene. The solid phase support is then contacted with a test nucleic acid and hybridization to the specific probes is detected. Accordingly, the identity of numerous allelic variants of one or more genes may be identified in a simple hybridization experiment.

**[181]** In certain embodiments, detection of the lesion comprises utilizing the probe/primer in a polymerase chain reaction (PCR) (*see, e.g.,* U.S. Patent Nos. 4,683,195 and 4,683,202), such as anchor PCR or RACE PCR, or, alternatively, in a ligase chain reaction (LCR) (*see, e.g.,* Landegran, et al., Science 241:1077-1080, 1988; Nakazaw, et al., Proc. Natl. Acad. Sci. USA 91:360-364, 1994), the latter of which can be particularly useful for detecting point mutations in the gene (*see, e.g.,* Abravaya, et al., Nuc. Acid Res. 23:675-682, 1995). In an illustrative embodiment, the method includes the steps of (i) collecting a sample of cells from a patient, (ii) isolating nucleic acid (e.g., genomic, mRNA, or both) from the cells of the sample, (iii) contacting the nucleic acid sample with one or more primers which specifically hybridize to a nucleic acid sequence under conditions such that hybridization and amplification of the nucleic acid (if present) occurs, and (iv) detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. It is anticipated that PCR and/or LCR may be desirable to use as a preliminary amplification step in conjunction with any of the techniques used for detecting mutations described herein.

**[182]** The invention thus, also encompasses methods of screening for agents which inhibit or enhance the expression of the nucleic acid markers *in vitro*, comprising exposing a cell or tissue in which the marker nucleic acid mRNA is detectable in cultured cells to an agent in order to determine whether the agent is capable of inhibiting or enhancing production of the mRNA; and determining the level of mRNA in the exposed cells or tissue, wherein a decrease in the level of the mRNA after exposure of the cell line to the agent is indicative of inhibition of the marker nucleic acid mRNA production and an increase in mRNA levels is indicative of enhancement of maker mRNA production.

**[183]** Alternatively, the screening method may include *in vitro* screening of a cell or tissue in which marker protein is detectable in cultured cells to an agent suspected of inhibiting or enhancing production of the marker protein; and determining the level of the marker protein in the cells or tissue, wherein a decrease in the level of marker protein after exposure of the cells or tissue to the agent is indicative of inhibition of marker protein production and an increase on the level of marker protein is indicative of enhancement of marker protein production.

**[184]** The invention also encompasses *in vivo* methods of screening for agents which inhibit or enhance expression of the marker nucleic acids, comprising exposing a subject having tumor cells in which marker mRNA or protein is detectable to an agent suspected of inhibiting or enhancing production of marker mRNA or protein; and determining the level of marker mRNA or protein in tumor cells of the exposed mammal. A decrease in the level of marker mRNA or protein after exposure of the subject to the agent is indicative of inhibition of marker nucleic acid expression and an increase in the level of marker mRNA or protein is indicative of enhancement of marker nucleic acid expression.

**[185]** Accordingly, the invention provides a method comprising incubating a cell expressing the marker nucleic acids with a test compound and measuring the mRNA or protein level. The invention further provides a method for quantitatively determining the level of expression of the marker nucleic acids in a cell population, and a method for determining whether an agent is capable of increasing or decreasing the level of expression of the marker nucleic acids in a cell population. The method for determining whether an agent is capable of increasing or decreasing the level of expression of the marker nucleic acids in a cell population comprises the steps of (a) preparing cell extracts from control and agent-treated cell populations, (b) isolating the marker polypeptides from the cell extracts, and (c) quantifying (e.g., in parallel) the amount of an immunocomplex formed between the marker polypeptide and an antibody specific to said polypeptide. The marker polypeptides of this invention may also be quantified by assaying for its bioactivity. Agents that induce an increase in the marker nucleic acid expression may be identified by their ability to increase the amount of immunocomplex formed in the treated cell as compared with the amount of the immunocomplex formed in the control cell. In a similar manner, agents that decrease expression of the marker nucleic acid may be identified by their ability to decrease the amount of the immunocomplex formed in the treated cell extract as compared to the control cell.

*Predictive Assays*

**[186]** Laboratory-based assays, which can predict clinical benefit from a given anti-cancer agent, will greatly enhance the clinical management of patients with cancer. In order to assess this effect, a biomarker associated with the anti-cancer agent may be analyzed in a biological sample (e.g., tumor sample, plasma) before, during, and following treatment.

**[187]** Another approach to monitor treatment is an evaluation of serum proteomic spectra. Specifically, plasma samples may be subjected to mass spectroscopy (e.g., surface-enhanced laser desorption and ionization) and a proteomic spectra may be generated for each patient. A set of spectra, derived from analysis of plasma from patients before and during treatment, may be analyzed by an iterative searching algorithm, which can identify a proteomic pattern that

completely discriminates the treated samples from the untreated samples. The resulting pattern may then be used to predict the clinical benefit following treatment.

**[188]** Global gene expression profiling of biological samples (e.g., tumor biopsy samples, blood samples) and bioinformatics-driven pattern identification may be utilized to predict clinical benefit and sensitivity, as well as development of resistance to an anti-cancer agent. For example, RNA isolated from cells derived from whole blood from patients before and during treatment may be used to generate blood cell gene expression profiles utilizing Affymetrix GeneChip technology and algorithms. These gene expression profiles may then predict the clinical benefit from treatment with a particular anti-cancer agent.

**[189]** Analysis of the biochemical composition of urine by 1D $^{1}$H-NMR (Nuclear Magnetic Resonance) may also be utilized as a predictive assay. Pattern recognition techniques may be used to evaluate the metabolic response to treatment with an anti-cancer agent and to correlate this response with clinical endpoints. The biochemical or endogenous metabolites excreted in urine have been well-characterized by proton NMR for normal subjects (Zuppi, et al., Clin Chim Acta 265:85-97, 1997). These metabolites (approximately 30-40) represent the by-products of the major metabolic pathways, such as the citric acid and urea cycles. Drug-, disease-, and genetic-stimuli have been shown to produce metabolic-specific changes in baseline urine profiles that are indicative of the timeline and magnitude of the metabolic response to the stimuli. These analyses are multi-variant and therefore use pattern recognition techniques to improve data interpretation. Urinary metabolic profiles may be correlated with clinical endpoints to determine the clinical benefit.

*Kits*

**[190]** The invention further provides kits for determining the expression level of genes characteristic of human breast tumor tissue. The kits may be useful for identifying subjects that are predisposed to developing cancer or who have cancer, as well as for identifying and validating therapeutics for cancer. In one embodiment, the kit comprises a computer readable medium on which is stored one or more gene expression profile of diseased cells of cancer, or at least values representing levels of expression of one or more genes characteristic of human breast tumor tissue in a diseased cell. The computer readable medium can also comprise gene expression profiles of counterpart normal cells, diseased cells treated with a drug, and any other gene expression profile described herein. The kit can comprise expression profile analysis software capable of being loaded into the memory of a computer system.

**[191]** A kit can comprise a microarray comprising probes of genes characteristic of human breast tumor tissue. A kit can comprise one or more probes or primers for detecting the expression level

of one or more genes characteristic of human breast tumor tissue and/or a solid support on which probes attached and which can be used for detecting expression of one or more genes characteristic of human breast tumor tissue in a sample. A kit may further comprise nucleic acid controls, buffers, and instructions for use.

**[192]** Other kits provide compositions for treating cancer. For example, a kit can also comprise one or more nucleic acids corresponding to one or more genes characteristic of human breast tumor tissue (e.g., for use in treating a patient having cancer). The nucleic acids can be included in a plasmid or a vector (e.g., a viral vector). Other kits comprise a polypeptide encoded by a gene characteristic of cancer or an antibody to a polypeptide. Yet other kits comprise compounds identified herein as agonists or antagonists of genes characteristic of human breast tumor tissue. The compositions may be pharmaceutical compositions comprising a pharmaceutically acceptable excipient.

## EXAMPLES

**[193]** It will be apparent to those skilled in the art that the examples and embodiments described herein are by way of illustration and not of limitation, and that other examples may be used without departing from the spirit and scope of the present invention, as set forth in the claims.

***Example 1. Gene Expression Profiling Protocol***

*A. Tissue Source*

**[194]** Human breast tumor tissue and normal adjacent tissue were purchased from the National Disease Research Institute.

*B. RNA extraction and cRNA preparation*

**[195]** Total RNA was extracted from the human tissues using TRIzol reagent (Life Technologies, MD) according to a modified vendor protocol which utilizes the RNeasy protocol (Qiagen, CA). After homogenization with a Brinkmann Polytron PT10/35 (Brinkmann, Switzerland) and phase separation with chloroform, samples were applied to RNeasy columns. RNA samples were treated with DNase I using RNase-free DNase Set (Qiagen, CA).

**[196]** After elution and quantitation with UV spectrophotometry, each sample was reverse transcribed into double-stranded cDNA using the Gibco Superscript II Choice System for RT-PCR according to vendor protocol (Invitrogen, CA).

**[197]** Samples were organically extracted and ethanol precipitated. Approximately 1 μg cDNA was then used in an *in vitro* transcription reaction incorporating biotinylated nucleotides using an RNA labeling kit (Enzo Diagnostics, NY). The resulting cRNA was put through RNeasy clean-up protocol and then quantified using UV spectrophotometry. The cRNA (15 μg) was fragmented in the presence of MgOAc and KOAc at 94°C. Fragmented RNA (10 μg) was loaded onto each array, one cRNA sample per array. Arrays were hybridized for 16 hours at 45°C rotating at 60 rpm in an Affymetrix GeneChip Hybridization Oven 640.

*C. Microarray Suite 5.0 analysis*

**[198]** Following hybridization, arrays were stained with Phycoerythrin-conjugated Streptavidin, placed in an Agilent GeneArray Scanner and then exposed to a 488 nm laser, causing excitation of the phycoerythrin. The Microarray Suite 5.0 software digitally converts the intensity of light given off by the array into a numeric value indicative of levels of gene expression. Because each array represents a single sample, tumor RNA was compared to the RNA isolated from normal adjacent tissue.

*D. Spotfire analysis*

**[199]** The purpose of this analysis is to generate sets of markers to distinguish between breast cancer and normal tissues (nucleic acid sequences SEQ ID NOs. 1-127 and corresponding amino acid sequences SEQ ID NOs. 128-254). Marker Set One (Table 1) represents a set of probe sets that is an optimum set for the prediction of whether or not a tissue is cancerous using a support vector machine. The optimal set is determined to be the one that shows the greatest prediction accuracy with the least error. This marker set was derived using the following method:

1. The data was imported into Spotfire.

2. A treatment comparison between cancer and normal tissues was performed using the t-test option.

3. The following criteria were used to select the probe sets:
    a. The data showed that the probe sets were all not "Absent," as determined by the Affymetrix Microarray Suite software v. 5.0 (Affymetrix, Inc., Santa Clara, CA) for either all of the normal or all of the cancer samples
    b. The data for the probe set showed a p-value of less than or equal to 0.001 according to the t-test.

4. All probe sets not meeting these criteria were eliminated from further analysis.

5. The remaining data was used in a selection process using custom software in conjunction with a modified version of the svm-train program (C++ version) which is part of LIBSVM (Chang and Lin, LIBSVM: A Library for Support Vector Machines, 2001. Software available from http://www.csie.ntu.edu.tw/~cjlin/libsvm). The custom software was written in the Perl language v. 5.004. The software was run on an SGI Origin 2000 running the IRIX 6.5.7f operation system. This software was used in the following manner:
    a. The Perl program was used as a "wrapper" to control svm-train. Its functions were to select subsets of the data and feed these sets to svm-train for training support vector machines (SVM).
    b. Training consisted of many elimination rounds. During each round many support vector machines were trained using ten-fold cross validation in order to estimate accuracy and error. Each SVM was trained on all data except that the data from one probe set was left out. One probe set was eliminated from

the data set at each round. This was the probe set that showed the best error and/or accuracy for the SVM when it was eliminated.

c. Training continued until there was only one probe set left.
d. The set of probe sets that showed the greatest accuracy with the least error was selected and is shown in Table 1.

The input arguments to svm-train were –s 0 –t 0 –c 1 –v 10.

**[200]** Marker Set Two (Table 2) represents a set of probe sets that is an optimum set for the prediction of whether or not a tissue is cancerous using a support vector machine. The optimal set is determined to be the one that shows the greatest prediction accuracy with the least error. This marker set was derived using the method described for Marker Set One with the following exceptions:

1. The data set was not limited to those probe sets that showed a t-test p-value of less than or equal to 0.001.
2. Five percent of the probe sets were eliminated at each round until 1000 probe sets remained. Then, only one probe set was eliminated during each round.

**[201]** Marker Set Three (Table 3) represents a set of probe sets that is an optimum set for the prediction of whether or not a tissue is cancerous using the Naïve Bayes method. The optimal set is determined to be the one that shows the greatest prediction accuracy. This marker set was derived using the following method:

1. The initial set of probe sets was the same as for Marker Set One.
2. The probe sets were then subjected to analysis with the Weka software version 3.3.4 (Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations Ian H. Witten, Eibe Frank., Morgan Kaufmann, October 1999)

    a. ForwardSelection search method was used to search for probe sets.
    b. Sets of probe sets were evaluated with the Naïve Bayes method.
    c. Ten-fold cross validation was used to select the eight most promising probe sets for further analysis.
    d. Probe sets were eliminated by manual manipulation until the optimal set remained which provided 100% prediction accuracy using the Naïve Bayes method.

**[202]** All three marker sets could select which tissues were cancerous and which were normal with 100% accuracy using their respective methods as determined by ten fold cross validation.

**Table 1. Breast Tumor Marker (Set One)**

| SEQ ID NO | Probe Set | Gene Symbol | Title | Genbank Accession | Unigene Cluster |
|---|---|---|---|---|---|
| 20 | 202021_x_at | SUI1 | putative translation initiation factor | AF083441.1 | Hs.150580 |
| 87 | 212914_at | PKP4 | plakophilin 4 | AV648364 | Hs.152151 |
| 55 | 208671_at | KIAA1253 | KIAA1253 protein | AF164794.1 | Hs.146668 |
| 102 | 215171_s_at | TIMM17A | translocase of inner mitochondrial membrane 17 homolog A (yeast) | AK023063.1 | Hs.20716 |
| 10 | 201488_x_at | KHDRBS1 | KH domain containing, RNA binding, signal transduction associated 1 | BC000717.1 | Hs.119537 |
| 32 | 203636_at | TRIM28 | tripartite motif-containing 28 | BE967532 | Hs.228059 |
| 104 | 215438_x_at | GSPT1 | G1 to S phase transition 1 | BE906054 | Hs.2707 |
| 1 | 200762_at | DPYSL2 | dihydropyrimidinase-like 2 | NM_001386.1 | Hs.173381 |
| 61 | 209307_at | SWAP70 | SWAP-70 protein | BC000616.1 | Hs.153026 |
| 89 | 213029_at | | Homo sapiens mRNA; cDNA DKFZp564H1916 (from clone DKFZp564H1916) | AL110126.1 | Hs.326416 |
| 15 | 201820_at | KRT5 | keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types) | NM_000424.1 | Hs.195850 |
| 118 | 219078_at | FLJ10252 | hypothetical protein FLJ10252 | NM_018040.1 | Hs.53913 |
| 13 | 201736_s_at | TEB4 | similar to S. cerevisiae SSM4 | BF000409 | Hs.380875 |
| 103 | 215281_x_at | POGZ | pogo transposable element with ZNF domain | AK000199.1 | Hs.107088 |
| 125 | 41856_at | | Homo sapiens mRNA; cDNA DKFZp586D0918 (from clone DKFZp586D0918) | AL049370 | Hs.13350 |
| 39 | 204135_at | DOC1 | downregulated in ovarian cancer 1 | NM_014890.1 | Hs.15432 |
| 17 | 201920_at | SLC20A1 | solute carrier family 20 (phosphate transporter), member 1 | NM_005415.2 | Hs.78452 |
| 75 | 212115_at | HN1L | HN1 like | AK023154.1 | Hs.172035 |
| 63 | 209479_at | DKFZP586D0623 | DKFZP586D0623 protein | BC000758.1 | Hs.44468 |
| 25 | 202373_s_at | RAB3-GAP150 | rab3 GTPase-activating protein, non-catalytic subunit (150kD) | AF255648.1 | Hs.197289 |
| 83 | 212596_s_at | THC211630 | partial transcript encompassing THC211630 gene | AL079310.1 | Hs.373557 |
| 119 | 219935_at | ADAMTS5 | a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 5 (aggrecanase-2) | NM_007038.1 | Hs.58324 |
| 57 | 208690_s_at | PDLIM1 | PDZ and LIM domain 1 (elfin) | BC000915.1 | Hs.75807 |
| 109 | 217795_s_at | MGC3222 | hypothetical protein MGC3222 | W74580 | Hs.323193 |
| 121 | 221729_at | COL5A2 | collagen, type V, alpha 2 | NM_000393.1 | Hs.82985 |
| 98 | 213893_x_at | PMS2L5 | postmeiotic segregation increased 2-like 5 | AA161026 | Hs.278895 |
| 79 | 212464_s_at | FN1 | fibronectin 1 | X02761.1 | Hs.287820 |
| 90 | 213165_at | | Homo sapiens clone CDABP0086 mRNA sequence | AI041204 | Hs.323748 |

| | | | | | |
|---|---|---|---|---|---|
| 68 | 209814_at | HSA6591 | nucleolar cysteine-rich protein | BC004421.1 | Hs.120766 |
| 24 | 202345_s_at | FABP5 | fatty acid binding protein 5 (psoriasis-associated) | NM_001444.1 | Hs.153179 |
| 110 | 217836_s_at | YAP | YY1 associated protein | NM_018253.1 | Hs.18851 |
| 50 | 205547_s_at | TAGLN | transgelin | NM_003186.2 | Hs.75777 |
| 2 | 200816_s_at | PAFAH1B1 | platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa | NM_000430.2 | Hs.77318 |
| 14 | 201737_s_at | TEB4 | similar to S. cerevisiae SSM4 | NM_005885.1 | Hs.20141 |
| 26 | 202404_s_at | COL1A2 | collagen, type I, alpha 2 | NM_000089.1 | Hs.179573 |
| 49 | 205158_at | RNASE4 | ribonuclease, RNase A family, 4 | NM_002937.1 | Hs.283749 |
| 11 | 201549_x_at | PLU-1 | putative DNA/chromatin binding motif | NM_006618.1 | Hs.143323 |
| 40 | 204141_at | TUBB | tubulin, beta polypeptide | NM_001069.1 | Hs.336780 |
| 48 | 205117_at | FGF1 | fibroblast growth factor 1 (acidic) | X59065 | Hs.75297 |
| 80 | 212573_at | KIAA0830 | KIAA0830 protein | AL573201 | Hs.167115 |
| 122 | 222151_s_at | FLJ13386 | hypothetical protein FLJ13386 | AK023738.1 | Hs.300876 |
| 9 | 201408_at | | ESTs | AI186712 | Hs.374442 |
| 93 | 213229_at | DICER1 | Dicer1, Dcr-1 homolog (Drosophila) | BF590131 | Hs.87889 |
| 30 | 203387_s_at | TBC1D4 | TBC1 domain family, member 4 | NM_014832.1 | Hs.173802 |
| 106 | 216336_x_at | | | AL031602 | |
| 62 | 209451_at | TANK | TRAF family member-associated NFKB activator | U59863.1 | Hs.146847 |
| 78 | 212227_x_at | SUI1 | putative translation initiation factor | AL516854 | Hs.150580 |
| 34 | 203640_at | MBLL39 | muscleblind-like protein MBLL39 | BE328496 | Hs.283609 |
| 127 | 823_at | CX3CL1 | chemokine (C-X3-C motif) ligand 1 | U84487 | Hs.80420 |
| 41 | 204204_at | SLC31A2 | solute carrier family 31 (copper transporters), member 2 | NM_001860.1 | Hs.24030 |
| 76 | 212130_x_at | SUI1 | putative translation initiation factor | AL537707 | Hs.150580 |
| 126 | 74694_s_at | FLJ23282 | hypothetical protein FLJ23282 | AA907940 | Hs.170253 |
| 18 | 201957_at | PPP1R12B | protein phosphatase 1, regulatory (inhibitor) subunit 12B | AF324888.1 | Hs.130760 |
| 101 | 214721_x_at | CDC42EP4 | CDC42 effector protein (Rho GTPase binding) 4 | AL162074.1 | Hs.3903 |
| 44 | 204688_at | SGCE | sarcoglycan, epsilon | NM_003919.1 | Hs.110708 |
| 27 | 202495_at | TBCC | tubulin-specific chaperone c | NM_003192.1 | Hs.75064 |
| 22 | 202225_at | CRK | v-crk sarcoma virus CT10 oncogene homolog (avian) | NM_016823.1 | Hs.343220 |
| 120 | 219956_at | GALNT6 | UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6) | NM_007210.2 | Hs.151678 |
| 52 | 208581_x_at | MT1X | metallothionein 1X | NM_005952.1 | Hs.374950 |
| 12 | 201648_at | JAK1 | Janus kinase 1 (a protein tyrosine kinase) | AL039831 | Hs.50651 |
| 23 | 202311_s_at | COL1A1 | collagen, type I, alpha 1 | NM_000088.1 | Hs.172928 |
| 69 | 210438_x_at | SSA2 | Sjogren syndrome antigen A2 (60kDa, ribonucleoprotein autoantigen SS-A/Ro) | M25077.1 | Hs.554 |

| | | | | | |
|---|---|---|---|---|---|
| 124 | 41329_at | LOC57147 | hypothetical protein LOC57147 | AI458463 | Hs.24243 |
| 81 | 212589_at | RRAS2 | related RAS viral (r-ras) oncogene homolog 2 | BG168858 | Hs.206097 |
| 8 | 201377_at | KIAA0144 | KIAA0144 gene product | NM_014847.1 | Hs.8127 |
| 113 | 218062_x_at | CDC42EP4 | CDC42 effector protein (Rho GTPase binding) 4 | NM_012121.2 | Hs.3903 |
| 72 | 211685_s_at | NCALD | neurocalcin delta | AF251061.1 | Hs.90063 |
| 107 | 216870_x_at | DLEU2 | deleted in lymphocytic leukemia, 2 | AF264787.1 | Hs.43628 |
| 95 | 213397_x_at | RNASE4 | ribonuclease, RNase A family, 4 | AI761728 | Hs.283749 |
| 16 | 201842_s_at | EFEMP1 | EGF-containing fibulin-like extracellular matrix protein 1 | AI826799 | Hs.76224 |
| 115 | 218823_s_at | FLJ20038 | hypothetical protein FLJ20038 | NM_017634.1 | Hs.72071 |
| 47 | 204976_s_at | AMMECR1 | Alport syndrome, mental retardation, midface hypoplasia and elliptocytosis chromosomal region, gene 1 | AK023637.1 | Hs.326142 |
| 85 | 212724_at | ARHE | ras homolog gene family, member E | BG054844 | Hs.6838 |
| 43 | 204619_s_at | CSPG2 | chondroitin sulfate proteoglycan 2 (versican) | BF590263 | Hs.81800 |
| 4 | 200961_at | SPS2 | selenophosphate synthetase 2 | NM_012248.1 | Hs.118725 |
| 29 | 203356_at | | Homo sapiens cDNA FLJ36423 fis, clone THYMU2011308 | BE349584 | Hs.377946 |
| 53 | 208662_s_at | TTC3 | tetratricopeptide repeat domain 3 | D84294.1 | Hs.118174 |
| 86 | 212730_at | DMN | desmuslin | AK026420.1 | Hs.10587 |
| 37 | 203973_s_at | CEBPD | CCAAT/enhancer binding protein (C/EBP), delta | NM_005195.1 | Hs.76722 |
| 21 | 202037_s_at | SFRP1 | secreted frizzled-related protein 1 | NM_003012.2 | Hs.7306 |
| 5 | 201012_at | ANXA1 | annexin A1 | NM_000700.1 | Hs.78225 |

**Table 2. Breast Tumor Marker (Set Two)**

| SEQ ID NO | Probe Set | Gene Symbol | Title | Genbank Accession | Unigene Cluster |
|---|---|---|---|---|---|
| 97 | 213647_at | DNA2L | DNA2 DNA replication helicase 2-like (yeast) | D42046.1 | Hs.194665 |
| 91 | 213180_s_at | GOSR2 | golgi SNAP receptor complex member 2 | BE730204 | Hs.100651 |
| 65 | 209635_at | AP1S1 | adaptor-related protein complex 1, sigma 1 subunit | BC003561.1 | Hs.57600 |
| 3 | 200925_at | COX6A1 | cytochrome c oxidase subunit VIa polypeptide 1 | NM_004373.1 | Hs.180714 |
| 33 | 203854_at | IF | I factor (complement) | NM_000204.1 | Hs.36602 |
| 96 | 213620_s_at | ICAM2 | intercellular adhesion molecule 2 | AA126728 | Hs.347326 |
| 59 | 209157_at | DNAJA2 | DnaJ (Hsp40) homolog, subfamily A, member 2 | AF011793.1 | Hs.21189 |
| 54 | 208663_s_at | TTC3 | tetratricopeptide repeat domain 3 | D84294.1 | Hs.118174 |
| 64 | 209621_s_at | ALP | alpha-actinin-2-associated LIM protein | AF002280.1 | Hs.135281 |
| 51 | 208158_s_at | OSBPL1A | oxysterol binding protein-like 1A | NM_018030.1 | Hs.252716 |
| 19 | 201989_s_at | CREBL2 | cAMP responsive element binding protein-like 2 | NM_001310.1 | Hs.13313 |
| 60 | 209268_at | VPS45A | vacuolar protein sorting 45A (yeast) | AF165513.1 | Hs.6650 |
| 105 | 215696_s_at | KIAA0310 | KIAA0310 gene product | BC001404.1 | Hs.5716 |
| 78 | 212227_x_at | SUI1 | putative translation initiation factor | AL516854 | Hs.150580 |
| 73 | 211896_s_at | DCN | decorin | AF138302.1 | Hs.76152 |
| 114 | 218226_s_at | NDUFB4 | NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 4, 15kDa | NM_004547.2 | Hs.227750 |
| 71 | 211285_s_at | UBE3A | ubiquitin protein ligase E3A (human papilloma virus E6-associated protein, Angelman syndrome) | U84404.1 | Hs.180686 |
| 77 | 212153_at | POGZ | pogo transposable element with ZNF domain | AB007930.1 | Hs.107088 |
| 84 | 212722_s_at | PSR | phosphatidylserine receptor | AK021780.1 | Hs.72660 |
| 6 | 201018_at | EIF1A | eukaryotic translation initiation factor 1A | BE542684 | Hs.4310 |
| 74 | 211961_s_at | RAB7 | RAB7, member RAS oncogene family | AK000826.1 | Hs.356386 |
| 20 | 202021_x_at | SUI1 | putative translation initiation factor | AF083441.1 | Hs.150580 |
| 36 | 203951_at | CNN1 | calponin 1, basic, smooth muscle | NM_001299.1 | Hs.21223 |
| 112 | 218039_at | ANKT | nucleolar protein ANKT | NM_016359.1 | Hs.279905 |
| 82 | 212590_at | RRAS2 | related RAS viral (r-ras) oncogene homolog 2 | BG168858 | Hs.206097 |
| 56 | 208684_at | COPA | coatomer protein complex, subunit alpha | U24105.1 | Hs.75887 |
| 67 | 209766_at | PRDX3 | peroxiredoxin 3 | AF118073.1 | Hs.75454 |
| 28 | 203213_at | CDC2 | cell division cycle 2, G1 to S and G2 to M | AL524035 | Hs.334562 |
| 99 | 214196_s_at | CLN2 | ceroid-lipofuscinosis, neuronal 2, late infantile (Jansky-Bielschowsky disease) | AA602532 | Hs.20478 |
| 45 | 204719_at | ABCA8 | ATP-binding cassette, sub-family A (ABC1), member 8 | NM_007168.1 | Hs.38095 |
| 70 | 210835_s_at | CTBP2 | C-terminal binding protein 2 | AF222711.1 | Hs.171391 |

| | | | | | |
|---|---|---|---|---|---|
| 100 | 214693_x_at | DJ328E19.C1.1 | hypothetical protein DJ328E19.C1.1 | BE732345 | Hs.218329 |
| 68 | 209814_at | HSA6591 | nucleolar cysteine-rich protein | BC004421.1 | Hs.120766 |
| 116 | 218930_s_at | FLJ11273 | hypothetical protein FLJ11273 | NM_018374.1 | Hs.3542 |
| 123 | 39582_at | | Homo sapiens mRNA; cDNA DKFZp586D1122 (from clone DKFZp586D1122) | AL050166 | Hs.26295 |
| 42 | 204257_at | FADS3 | fatty acid desaturase 3 | NM_021727.1 | Hs.21765 |
| 46 | 204905_s_at | EEF1E1 | eukaryotic translation elongation factor 1 epsilon 1 | NM_004280.1 | Hs.298581 |
| 108 | 217106_x_at | HSA9761 | putative dimethyladenosine transferase | AF091078.1 | Hs.125819 |
| 92 | 213225_at | PPM1B | protein phosphatase 1B (formerly 2C), magnesium-dependent, beta isoform | AJ271832.1 | Hs.5687 |
| 66 | 209687_at | CXCL12 | chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1) | U19495.1 | Hs.237356 |
| 117 | 219064_at | MGC10848 | hypothetical protein MGC10848 | NM_030569.1 | Hs.207443 |
| 7 | 201143_s_at | EIF2S1 | eukaryotic translation initiation factor 2, subunit 1 alpha, 35kDa | BC002513.1 | Hs.151777 |
| 94 | 213361_at | PCTAIRE2BP | tudor repeat associator with PCTAIRE 2 | AW129593 | Hs.283761 |
| 119 | 219935_at | ADAMTS5 | a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 5 (aggrecanase-2) | NM_007038.1 | Hs.58324 |
| 121 | 221729_at | COL5A2 | collagen, type V, alpha 2 | NM_000393.1 | Hs.82985 |
| 111 | 217901_at | | Homo sapiens, clone IMAGE:4242700, mRNA | BF031829 | Hs.348710 |
| 38 | 204017_at | KDELR3 | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 | NM_006855.2 | Hs.250696 |
| 61 | 209307_at | SWAP70 | SWAP-70 protein | BC000616.1 | Hs.153026 |
| 58 | 208944_at | TGFBR2 | transforming growth factor, beta receptor II (70/80kDa) | D50683.1 | Hs.82028 |
| 35 | 203810_at | DNAJB4 | DnaJ (Hsp40) homolog, subfamily B, member 4 | BG252490 | Hs.41693 |
| 31 | 203567_s_at | TRIM38 | tripartite motif-containing 38 | NM_006355.1 | Hs.59545 |
| 88 | 213006_at | CEBPD | CCAAT/enhancer binding protein (C/EBP), delta | AV655640 | Hs.76722 |
| 9 | 201408_at | | ESTs | AI186712 | Hs.374442 |
| 37 | 203973_s_at | CEBPD | CCAAT/enhancer binding protein (C/EBP), delta | NM_005195.1 | Hs.76722 |
| 126 | 74694_s_at | FLJ23282 | hypothetical protein FLJ23282 | AA907940 | Hs.170253 |
| 29 | 203356_at | | Homo sapiens cDNA FLJ36423 fis, clone THYMU2011308 | BE349584 | Hs.377946 |
| 5 | 201012_at | ANXA1 | annexin A1 | NM_000700.1 | Hs.78225 |

**Table 3. Breast Tumor Marker (Set Three)**

| SEQ ID NO | Probe Set | Gene Symbol | Title | Genbank Accession | Unigene Cluster |
|---|---|---|---|---|---|
| | | | | | |
| 1 | 200762_at | DPYSL2 | dihydropyrimidinase-like 2 | NM_001386.1 | Hs.173381 |
| 44 | 204688_at | SGCE | sarcoglycan, epsilon | NM_003919.1 | Hs.110708 |